Alimerkkijonot suomen sanojen vektoriesitysten tuottamisessa neuroverkoilla

Ada-Maaria Hyvärinen

Pro gradu
HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen osasto

Helsinki, 15. lokakuuta 2018

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Matemaattis-luonnontieteellinen	Tietojenkäsittelytieteen osasto

Tekijä — Författare — Author
Ada-Maaria Hyvärinen

Työn nimi — Arbetets titel — Title

Alimerkkijonot suomen sanojen vektoriesitysten tuottamisessa neuroverkoilla

Oppiaine — Läroämne — Subject
Tietojenkäsittelytiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu	15. lokakuuta 2018	53

Tiivistelmä — Referat — Abstract

Sanojen vektoriesityksiä käytetään moniin luonnollista kieltä käsitteleviin koneoppimistehtäviin, kuten luokitteluun, tiedonhakuun ja konekääntämiseen. Ne ilmaisevat sanat tietokoneelle ymmärrettävässä muodossa. Erityisen hyödyllinen tapa esittää sanat vektoreina on esittää sanat pisteinä jatkuvassa sana-avaruudessa, jolla on joitakin satoja ulottuvuuksia. Tällaisessa mallissa samankaltaiset sanat sijaitsevat avaruudessa lähekkäin, ja sanavektorien erotukset kuvaavat sana-analogiasuhteita, jos vektorit on tuotettu siihen tarkoitukseen luodulla neuroverkolla. Pelkästään tällaisia vektoreita katsomalla saadaan tietää jotakin sanan merkityksestä ja muodosta.

Perinteisesti sanavektoreita opettaessa on käsitelty opetusaineiston sanat erillisinä merkkijonoina. Englannin kielessä tämä on usein toimiva menetelmä. Suomen kieli taas on vahvasti taivuttava, joten myös sananmuodot sisältävät paljon informaatiota. Osa informaatiosta menee hukkaan, jos sanat opetetaan kokonaan erillisinä. Lisäksi malli ei osaa yhdistää kahta saman sanan sanamuotoa toisiinsa.

FastText-mallit ratkaisevat taivuttamisen ja johtamisen tuomat ongelmat hyödyntämällä tietoa sanojen sisältämistä alimerkkijonoista. Vektoriesitysmalli opetetaan siis paitsi sanojen, myös niiden sisältämien lyhyempien merkkijonojen perusteella. Tämän takia fastText-mallin voisi ajatella toimivan hyvin paljon taivuttavilla kielillä, kuten suomella.

Tässä tutkielmassa on haluttu selvittää, toimiiko fastText-menetelmä hyvin suomen kielellä. Lisäksi on tutkittu, millä parametreilla malli toimii parhaiten. Tutkielmassa on kokeiltu erilaisia alimerkkijonojen pituuksia ja sanavektorin kokoja.

Mallin laatua voidaan testata semanttista samankaltaisuutta mittaavilla aineistoilla sekä sana-analogiakyselyillä. Semanttista samankaltaisuutta mittaavissa testeissä tutkitaan, ovatko samaa tarkoittavat sanat lähekkäin vektoriavaruudessa. Aineistot pohjautuvat ihmisarvioijien antamiin pisteytyksiin sanojen samankaltaisuudesta.

Sana-analogiatesteissä kokeillaan, onnistuuko malli löytämään analogiaparista puuttuvan sanan vektorilaskutoimituksen perusteella. Analogia-aineistot koostuvat sanapareista, jotka ovat tietyssä analogiasuhteessa keskenään. Analogiat voivat liittyä sanan merkitykseen, kuten “mies ja nainen” tai muotoon, kuten “positiivi ja komparatiivi”.

Tutkielmaa varten käännettiin suomeksi kaksi englannin kielellä usein käytettyä datasettiä: semanttista samankaltaisuutta mittaava WS353 ja sana-analogioita sisältävä SSWR, jonka käännöksestä käytetään nimeä SSWR-fi. Käännöksissä huomioitiin se, että monet datasettien sanat eivät käänny suomeen yksikäsitteisesti. SSWR-fi-datasetistä ongelmalliset sanat poistettiin, WS353-datasetin rinnalle taas tehtiin erillinen lyhennetty datasetti WS277-josta ongelmalliset sanat on poistettu.

Tutkielmassa havaittiin, että alimerkkijonojen käyttäminen on hyödyllistä suomen kielen käsittelyssä. Semanttista samankaltaisuutta mittaavien testien mukaan mallin laatu parani alimerkkijonojen ansiosta. Sana-analogiatesteissä alimerkkijonojen käyttäminen paransi muotokyselyissä onnistumista, mutta huononsi merkityskyselyissä onnistumista. Tämä johtunee siitä, että muotokyselyt perustuvat sanojen taivuttamiselle ja johtamiselle, mutta merkityskyselyissä sananmuodoilla ei ole juuri väliä.

Avainsanat — Nyckelord — Keywords
luonnollisen kielen käsittely, koneoppiminen, vektoriesitys

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Sisältö

1 Johdanto **1**
1.1 Termistöstä . . . 3

2 Sanojen vektoriesitysmallit **4**
2.1 Word embedding -mallit . . . 5
2.2 Word2vec . . . 6
2.2.1 Yksinkertaistettu versio mallista . . . 8
2.2.2 CBOW ja skip-gram . . . 9
2.3 fastText-malli . . . 10
2.3.1 Negatiivista näytteistystä käyttävä skip-gram-malli . . . 11
2.3.2 Alimerkkijonojen hyödyntäminen . . . 12
2.3.3 Mallin toteutuksen yksityiskohtia . . . 13
2.4 Mallin opettaminen . . . 13
2.4.1 Esiprosessointi . . . 13

3 Semanttinen samankaltaisuus **14**
3.1 Datasetit . . . 14
3.1.1 WS353 ja WS277 . . . 16
3.1.2 SimLex . . . 18
3.1.3 FinnSim . . . 19
3.2 Tulokset . . . 19
3.2.1 Vektorin ulottuvuuksien määrä . . . 20
3.2.2 N-grammien pituus . . . 21
3.2.3 Tulosten vertailua . . . 24

4 Analogiatestit **26**
4.1 Datasetit . . . 27
4.1.1 SSWR ja SSWR-fi . . . 28
4.1.2 FinnAn . . . 30
4.2 Tulokset . . . 32
4.2.1 Vektorin ulottuvuuksien määrä . . . 32
4.2.2 N-grammien pituus . . . 35
4.2.3 Tulosten vertailua . . . 40
4.2.4 Mistä ero muoto- ja merkityskyselyiden välillä voisi johtua? . . . 41
4.2.5 Havaintoja yleisistä virheistä . . . 42

5 Johtopäätökset ja tuleva tutkimus **45**
5.1 Onko alimerkkijonojen käyttö suomen kielellä hyödyllistä? . . . 45
5.2 Mitkä parametrit toimivat parhaiten? . . . 46
5.3 Tuleva tutkimus . . . 47

Lähteet **48**

1 Johdanto

Yksi luonnollisen kielen käsittelyn perusyksiköistä on sana. Kieliteknologian sovelluksissa tietoa sanoista tarvitaan monenlaisiin tehtäviin: dokumenttien luokitteluun [21][9], tekstin tuottamiseen [25][43], keskusteleviin tietokoneohjelmiin [44][37] ja konekääntämiseen [5][45].

Kun sanoja käsitellään sääntöpohjaisilla menetelmillä, voi ne usein esittää merkkijonoina. Koneoppimista varten tarvitaan kuitenkin tieto sanoista numeerisessa muodossa. Tästä syntyy motivaatio esittää sanat vektoreina. Sanojen vektoriesitys on siis yleisimmissä muodossaan mikä tahansa tapa kuvata aineiston sanat vektoreina. Tässä tutkielmassa keskitytään kuitenkin esityksiin, joissa sanavektori koostuu liukuluvuista ja on tuotettu neuroverkon avulla.

Tällaisilla sanavektoreilla on mahdollista tehdä paljon muitakin vertailuja kuin pelkkä yhtäsuuruusvertailu. Hyvällä mallilla sanavektoreilla on samankaltaisuusominaisuus, eli samankaltaisten sanojen vektorit ovat lähekkäin ja erilaisten sanojen kaukana toisistaan. Lisäksi kahden sanan suhdetta voidaan kuvata sanavektorien erotuksella.

Suomen kieltä on pidetty perinteisesti vaikeana ymmärtää tietokoneille, sillä suomi on paljon agglutinoiva eli taivuttava ja morfologisesti kompleksinen kieli, eli säännöt suomen kielen sananmuotojen muodostamiseen ovat monimutkaiset. Lisäksi suomessa voidaan muodostaa uusia yhdyssanoja tunnettujen sanojen perusteella. Kun aineisto jaetaan saneisiin välilyöntien ja -merkkien perusteella, on merkityksellisiä sananmuotoja lukumäärällisesti vähemmän englannin- kuin suomenkielisellä aineistolla. Ilman perusmuodottamista suomen kielen sananmuodot *suomi* ja *suomen* näyttäytyvät koneoppimismallille toisistaan riippumattomina, vaikka suomea osaava puhuja osaa heti sanoa niiden olevan saman lekseemin kaksi eri sananmuotoa. Lisäksi sanojen yhdistämisestä seuraavat uudissanat jäävät koneoppimismalleille tuntemattomiksi: englanninkielisellä aineistolla koneoppimismalli tunnistaisi esikäsitellystä aineistosta uutuusleivonnaisen *cactus pie* liittyvän kaktuksiin ja piirakoihin. Sen sijaan *kaktuspiirakka* käsiteltäisiin sanastoon kuulumattomana sanana, koska mallin pitäisi ensin ymmärtää sanan olevan yhdyssana ja sitten osata jakaa se osiin.

Perinteisesti ongelmaa on lähestytty kieliriippuvaisilla työkaluilla, jotka osaavat palauttaa sananmuodon perusmuotoon ja kertoa, missä muodossa sana on (esim. Pirinen [30]). Toisaalta myös kieliriippumattomia morfeemien eli sanan merkitysyksikköjen tunnistamiseen tarkoitettuja menetelmiä on kehitetty, esimerkkinä Creutzin ym. Morfessor [8]. Työkaluilla on mahdollista myös muodostaa sopiva jako yhdyssanan eri osille. Kaikille kielille tällaisia työkaluja ei kuitenkaan ole saatavilla, tai ne voivat toimia huonosti alikielillä, kuten suomen puhekielellä ja murteilla. Lisäksi ilman tietoa sanan kontekstista voi olla mahdotonta tietää, onko sana yhdyssana ja minkä sanan taipunut muoto se on. Esimerkiksi *piilevät* voisi olla joko muoto *piillä*-verbistä tai

piilevä-kasvista.

Yksi tapa vastata tähän ongelmaan on käyttää alimerkkijonoja, kuten fastText-neuroverkkomallit tekevät. Mallit kuvataan täsmällisesti luvussa 2.3. Alimerkkijonoja voidaan käyttää ilman mitään tietoa kielen rakenteesta, mutta siitä huolimatta ne antavat mahdollisuuden jakaa parametreja samoja alimerkkijonoja sisältävien sanojen välillä. Bojanowski ym. [4] esittävät, että alimerkkijonojen käyttö voisi parantaa kielimallien laatua morfologisesti rikkailla kielillä, kuten suomella. He eivät kuitenkaan tutkineet mallien toimintaa suomella. Tässä tutkielmassa pyritään selvittämään, pitääkö hypoteesi paikkansa.

Tutkielman tarkoituksena on arvioida erilaisten fastText-mallien toimintaa ja tutkia, parantaako alimerkkijonojen käyttö suomenkielisten sanojen vektoriesitysten laatua. Toiminnan arvioinnissa on kaksi tavoitetta. Toinen tavoite on vertailla erilaisten mallien suoriutumista semanttista samankaltaisuutta ja sanojen analogisuutta mittaavista teksteistä. Tarkoituksena on saada tietoa siitä, millaisilla parametreilla malli kannattaa opettaa, jos siitä haluaa saada mahdollisimman laadukkaan.

Toinen tavoite on ollut verrata tutkielmaa varten opetettujen mallien laatua muussa kirjallisuudessa esitettyjen mallien laatuun. Aiemmin suomenkielisiä vektoriesitysmalleja ovat testanneet Venekoski ym. [42], joiden mukaan eri parametrien kokeilu voisi kasvattaa mallin laatua paljonkin.

Suomenkielisiä malleja voidaan myös verrata muunkielisiin malleihin. Usein muilla kuin englannin kielellä tehdyissä kokeissa saadaan huonompia tuloksia, vaikka testaamiseen käytettäisiin samaa datasettiä tutkittavalle kielelle käännettynä. Tämä näkyy myös Bojanowskin ym. [4] tekemissä kokeissa, joissa vertailtiin fastText-mallien toimintaa esimerkiksi saksan- ja arabiankielisillä aineistoilla. Tässä tutkielmassa selvitetään, pääseekö suomenkielinen malli lähemmäs englanninkielisten mallien tasoa alimerkkijonojen ansiosta. Jos näin käy, voidaan sanoa, että alimerkkijonoista on hyötyä suomenkielisten mallien opettamisessa. Vertailuun liittyy kuitenkin monia ongelmia, jotka aiheutuvat datasettien kääntämisen tarpeesta sekä siitä, että suomen ja englannin opetusaineistot eivät ole samoja.

Tutkielman rakenne on seuraava: Ensimmäisessä luvussa esitellään tutkielman aihe ja avataan tutkielmassa käytettyä termistöä. Toisessa luvussa kerrotaan sanojen vektoriesitysmalleista. Siinä esitellään word2vec-mallit ja niitä laajentava fastText-malli, jota tässä tutkielmassa on tarkasteltu. Kolmannessa luvussa tutkitaan, miten hyvin eri parametreilla opetetut fastText-mallit selviytyvät semanttista samankaltaisuutta mittaavista kokeista. Neljännessä luvussa nähdään, miten mallit pärjäävät sana-analogioiden löytämistä arvioivissa testeissä. Viimeisessä luvussa käydään läpi tutkimuksesta syntyneet johtopäätökset ja pohditaan tulevia tutkimussuuntia.

1.1 Termistöstä

Tämä on tietojenkäsittelytieteen alan tutkielma, joten suurin osa siinä esiintyvästä termistöstä on tietojenkäsittelytieteen käytäntöjen mukaista. Aihe kuitenkin liittyy kiinteästi myös kieliteknologiaan, joten mahdollisuuksien mukaan tutkielmassa on pyritty käyttämään myös kieliteknologian käytäntöjen mukaista termistöä.

Tietojenkäsittelytieteen alan julkaisuissa voidaan käsitteellä **sana** (word) viitata kielelliseen yksikköön. Kieliteknologian tarkoituksiin **sana** on usein liian epätarkka termi, vaan yleensä on tärkeää erotella, tarkoitetaanko **sanetta**, **lekseemiä** vai **sananmuotoa**. Tässä tutkielmassa erottelu ei useinkaan ole tärkeää, sillä **sanalla** tarkoitetaan aineistossa esiintynyttä välilyönneillä ja yhdysmerkeillä toisista sanoista eroteltua yksikköä, jonka merkityksestä voidaan puhua ja joka voi olla taipuneessa muodossa.

Joissain tilanteissa on kuitenkin tärkeää ottaa kantaa siihen, mitä **sanalla** täsmällisesti tarkoitetaan. **Lekseemi** (lexeme) on perusmuotoinen sana, jollaista voi käyttää vaikkapa sanakirjan hakusanana: lekseemejä ovat esimerkiksi *hyvä*, *kampela* tai *litistyä*. **Sananmuoto** (word form) on sana taipuneessa muodossa, joka voi olla myös samannäköinen kuin sen perusmuoto, eli esimerkiksi *asioitahan*, *kääntää* ja *muita* ovat sananmuotoja. **Sane** (token) on sananmuodon esiintymä aineistossa. Esimerkiksi lauseessa *oi, sanoi siili, olen surullinen siili* on 6 sanetta, ja sane *siili* esiintyy kahdesti. Monesti tässä tutkielmassa termillä **sana** viitataan nimenomaan **saneeseen**.

Vaikka tämän julkaisun kieli ei ole kauttaaltaan kieliteknologian käytäntöjen mukaista, kieliteknologisesta näkökulmasta mahdollisesti harhaanjohtavaa termistöä on pyritty kuitenkin välttämään. Tietojenkäsittelytieteen alan julkaisuissa puhutaan toisinaan sanojen **syntaksista** (syntax) ja **semantiikasta** (semantics), missä semantiikalla tarkoitetaan sanan merkitystä ja syntaksilla kaikkea sananmuotoihin ja sanan rakenteeseen liittyvää. Semanttista tietoa on esimerkiksi tieto siitä, että sana *kuusi* on yhdeltä merkitykseltään havupuu, mutta merkitsee myös viittä suurempaa luonnollista lukua sekä yksikön toisen persoonan omistamaa taivaankappaletta. Tietojenkäsittelytieteen alalla syntaktiseksi tiedoksi sanotaan tietoa siitä, että *taivaankappale*-merkityksessä sana muodostuu osista *kuu* ja omistusliite *si*. Kielitieteessä tätä kutsutaan **morfologiaksi** (morphology), ei syntaksiksi.

Kielitieteessä syntaksilla viitataan useimmiten lauseen syntaksiin eli siihen, missä roolissa sanat esiintyvät lauseessa ja millaisia suhteita niillä on toisiinsa. Tällaista tietoa on esimerkiksi tieto siitä, että lauseessa *pihalla kasvaa kaunis puu* lauseen predikaatti on *kasvaa* ja että *kaunis puu* on nominaalilauseke, jossa adjektiivi *kaunis* määrittää substantiivia *puu*. Tietojenkäsittelyssä tällaisesta lausetason informaatiosta puhutaan toisinaan sanojen välisinä **dependensseinä**.

Koska tämän tutkielman tulokset voivat olla kiinnostavia sekä tietojenkäsittelytieteen että kielitieteen näkökulmasta, tutkielmassa on pyritty

välttämään mahdollisesti sekaannuksia aiheuttavaa terminologiaa. Tutkielmassa puhutaan siis syntaksin sijaan morfologiasta, kun aiheena on sanojen merkitystä kantavat osaset. Silti esimerkiksi dataseteistä puhuttaessa on säilytetty osa alkuperäisten julkaisijoiden terminologiasta. Esimerkkinä suhteesta **vastakohta** on puhuttu kirjallisuudessa syntaktisena ominaisuutena, jos vastakohdat voidaan muodostaa tekemällä johdannainen sanasta, vaikka vastakohtaisuus onkin semanttinen ominaisuus. Esimerkiksi *epäeettinen* on sanan *eettinen* vastakohta, joka on muodostettu lisäämällä vastakohtaisuutta ilmaiseva etuliite *epä*. Tällöin tutkielmassa on puhuttu johdannaisista erotuksena sellaisista semanttisista suhteista, joita ei voida tuottaa kantasanan perusteella: esimerkiksi *pieni* on sanan *suuri* vastakohta, eikä sanaa *epäpieni* ole olemassa.

2 Sanojen vektoriesitysmallit

Sanojen vektoriesitysmalleilla tarkoitetaan malleja, joissa sana on esitetty vektoriavaruuden pisteenä.

Yksinkertainen tapa esittää sanat vektoreina on niin sanottu *one-hot-encoding*. Siinä vektorin pituus on sanaston koko, käytännössä esimerkiksi 10 000. Jokaista sanaston sanaa vastaa vektori, joka on sanaa kuvaavan indeksin kohdalla 1, ja muualla 0. Jokaista sanaa kuvaava indeksi on eri. Tällaisten vektorien perusteella voidaan vertailla, ovatko sanat samat. Sanoista ei kuitenkaan saada mitään muuta tietoa vektorien perusteella, sillä ne ovat kaikki ortogonaalisia toisiinsa nähden.

Word embedding -malleissa sanavektorit eivät ole binäärisiä vaan jatkuva-arvoisia vektoreita. Jokainen vektorin alkio on liukuluku, joka ilmaisee, kuinka voimakkaasti vektori on painottunut komponentin mukaisesti. Yksi tapa saada intuitiota word embedding -malleihin on kuvitella, että jokainen indeksi kuvaa jotakin sanan ominaisuutta: vaikkapa sana *hiiri* voisi olla painottunut indeksien **pienikokoisuus** ja **eläimellisyys** osalta, sen sijaan *norsu* olisi painottunut vain indeksissä **eläimellisyys** ja saisi matalan arvon indeksille **pienikokoisuus**. Todellisuudessa indeksien kuvaamat ominaisuudet eivät kuitenkaan ole suoraan tulkittavissa sanan merkityksen tai muodon kannalta. Ominaisuudet eivät myöskään ole ennalta määrättyjä: siinä missä one-hot-encoding-menetelmällä jokainen indeksi vastaa tiettyä sanaa, word embedding -mallien sanavektorit opitaan neuroverkon avulla ilman tiettyä tulkintaa jokaiselle indeksille.

Tässä tutkielmassa ollaan erityisen kiinnostuneita alimerkkijonoja hyödyntävistä fastText-malleista, jotka pohjautuvat word2vec-malleihin. Yksi ensimmäisestä word embedding -malleista oli Bengion ym. [2] esittelemä, mutta tehokkaamman word2vec-mallin esittelivät Mikolov ym. [26].

2.1 Word embedding -mallit

Jakaumahypoteesin (distributional hypothesis) mukaan "sanan luonne määrittyy sen seuralaisten mukaan" (Firth [11]). Word embedding -malleissa hyödynnetään tätä ajatusta olettamalla, että sanan merkitys voidaan päätellä sitä ympäröivistä sanoista. Jos siis tehdään tilastollinen malli suuresta määrästä sanoja, voidaan kahta sanaa vertailla tutkimalla niiden kontekstien samanlaisuutta. Hypoteesia tukevat ihmisarvioijilla tehdyt kokeet, joissa arvioijilta on kysytty arviota sanojen samankaltaisuudesta joko pelkkien sanaparien avulla (Rubenstein [34], Miller ym. [28]) tai näytetty heille sanoja erilaisissa konteksteissa (Charles [6]). Voidaankin sanoa, että ihmiset hallitsevat sanat *kontekstiesityksen*, eikä sanakirjamääritelmän perusteella: he tietävät, millaisissa tilanteissa sana on tuotettava tai miten se on ymmärrettävä kieltä käytettäessä [28]. Voidaan myös sanoa, että tällä tavalla jakaumaan perustuvat mallit jäljittelevät ihmisten kielenkäyttöä läheisemmin kuin sellaiset mallit, jotka perustuvat sanastoon tai ontologioihin eli sanojen keskinäisiin riippuvuussuhteisiin (esim. *koira* kuuluu luokkaan *eläin*) [33].

Word embedding -malleissa sanat esitetään pisteinä jatkuvassa vektoriavaruudessa. Esitys poikkeaa esimerkiksi ontologioihin perustuvien mallien luomista esityksistä siinä, että samankaltaiset sanat sijaitsevat lähellä toisiaan, ja että malli huomioi monia erilaisia samankaltaisuuden asteita [26]. Samankaltaisuus ei ole käsitteenä hyvin määritelty, vaan se voi tarkoittaa semanttista samankaltaisuutta, samaa kantaa olemista eli johdoksellista samankaltaisuutta, morfologista samankaltaisuutta eli samanlaisten taivutusmuotojen sisältämistä sekä samaan aihepiiriin kuulumista. Voidaan havaita, että sanat sijaitsevat avaruudessa lähekkäin, jos ne ovat samassa sanamuodossa (*Helsingissä*, *bussissa*), liittyvät samaan aiheeseen (*Helsinki*, *Pariisi*), ovat samaa kantaa (*johdonmukainen*, *epäjohdonmukainen*) tai merkitsevät suurin piirtein samaa asiaa (*kaupunki*, *kylä*).

Lähekkäisyyden lisäksi vektoriesitysmalleissa voidaan huomata sanojen välisiä suhteita, jotka esiintyvät samankaltaisina sellaisilla sanapareilla, joiden suhde on samankaltainen. Esimerkiksi sanapareilla *Pariisi*, *Ranska* ja *Berliini, Saksa*, on suhde **maan pääkaupunki**. Onnistuneessa sana-avaruudessa pisteiden *Pariisi*, *Ranska* suhde on samanlainen kuin pisteiden *Berliini*, *Saksa*. Voidaan siis tutkia sanojen suhteita toisiin sanoihin sen perusteella, millainen erotus sanavektoreilla on. Sanoilla voi olla paitsi merkitykseen liittyvä yhteys, myös muotoon liittyvä yhteys, eli sanaparin *Helsinki, Helsingissä* ja *bussi, bussissa* keskinäiset suhteet ovat samankaltaiset. Vastaava suhde on myös *Tampere, Tampereella*, eli sananmuodon ei välttämättä tarvitse olla sama, sillä sijaitsemista ei ilmaista lekseemin *Tampere* kohdalla inessiivillä vaan adessiivilla.

Vaikka verkkoa opetettaessa yksittäisille neuroneille ei anneta tulkittavissa olevaa merkitystä, opetuksen tuloksena sellainen voi kuitenkin syntyä. Koska vektoreiden erotukset kuvaavat sanojen välisiä suhteita, voi käydä ilmi,

että esimerkiksi **maskuliinisuus** esiintyy jonkinlaisena lineaarikombinaationa sana-avaruuden ulottuvuuksista. On mahdollista, että lineaarikombinaatio on sattumalta jonkin avaruuden ulottuvuuden suuntainen, jolloin ulottuvuudelle löytyisikin tosimaailman semanttinen tai morfologinen tulkinta. Tämän tutkielman kannalta sellainen on epätodennäköistä, sillä tutkielmassa käytetyt word2vec-pohjaiset mallit ovat varsin yksinkertaisia. Monimutkaisemmilla menetelmillä luotujen kielimallien sanavektoreilla on kuitenkin havaittu kokeellisesti myös suoraan tulkittavissa semanttisia ominaisuuksia, kuten Radfordin ym. [32] multiplikatiivinen LSTM-verkko, jossa yksi mallin 4096 yksiköstä havaittiin olevan vastuussa kohteena olevan saneen sävystä (sentiment).

Vektoriesitysmallien laatua voidaan mitata testeillä, joissa tutkitaan mallien kykyä laittaa samanlaiset sanat lähekkäin. Toinen tapa testata mallia on tutkia, vastaavatko sanavektorien väliset suhteet sanaparien välisiä suhteita, kun vertaillaan niitä toisiin saman suhteen sanapareihin. Muitakin mahdollisuuksia mallien testaamiseen löytyy. Usein vektoriesitysmalleja tarvitaan johonkin käytännön tehtävään, jolloin voidaan tutkia sitä, miten hyvin ne soveltuvat haluttuun tarkoitukseen. Esimerkiksi jos tavoitteena on käyttää sanavektoreita konekääntämisessä, on tarpeen tutkia käännöksen laatua erilaisilla malleilla, jotta voidaan valita, mitä halutaan käyttää.

Sanojen esittäminen merkitykseen liittyvinä sijainteina ei ole ennenkuulumatonta, vaan sitä tehdään myös kielitieteen puolella. Kielitieteessä samoissa konteksteissa esiintyvien sanojen kohdalla voidaan puhua **kollokaatiosta**. Sillä tarkoitetaan samaan merkityskenttään kuuluvia sanoja, joilla on luontainen taipumus esiintyä yhdessä [20]. Kollokaation käsitettä ja sanan ominaisuuksien mieltämistä vektoriavaruuden ulottuvuuksina on tehty kauan myös tiedonhaun ja informaatiotutkimuksen piirissä esimerkiksi luomalla tilastoja siitä, mitkä sanat esiintyvät usein yhdessä [36]. Word embedding -mallit rakentavat tämän tutkimuksen pohjalle.

2.2 Word2vec

Tässä tutkielmassa on hyödynnetty word2vec-malleihin pohjautuvia word embedding -menetelmiä. Word2vec-mallissa sanojen vektoriesitykset opetetaan yksinkertaisen neuroverkon avulla suuresta määrästä dataa. Ajatuksena on, että samanlaisissa konteksteissa esiintyvät sanat saavat toisiaan lähellä olevat vektoriesitykset, koska niiden merkitys on todennäköisesti samankaltainen. Siksi vektoriesityksiä opetettaessa hyödynnetään tietoa sanan kontekstista.

Kontekstisanat ovat sanoja, jotka esiintyvät tekstissä enintään etäisyyden c päässä kohdesanasta. Esimerkiksi lauseessa *hän kiertelee aihetta kuin kissa kuumaa puuroa* sanan *kissa* kontekstisanat ovat $\{$*aihetta*, *kuin*, *kuumaa*, *puuroa*$\}$, jos $c = 2$. Yleensä käytetään konteksti-ikkunaa, jossa c on melko pieni.

Mikolov ym. [26] käyttivät menetelmää, jossa c:lle arvotaan satunnainen

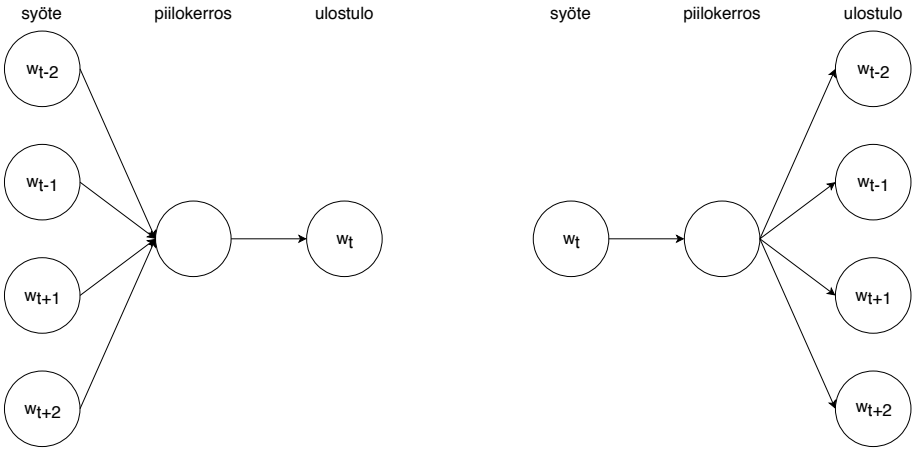
koko, joka vaihtelee välillä $[1, 10]$. Ajatuksena on, että on vaikea sanoa, mikä konteksti-ikkunan koon tulisi täsmälleen olla, eli millä etäisyydellä kohdesanan kannalta oleelliset sanat siitä aina sijaitsevat. Vaihtelemalla konteksti-ikkunan kokoa satunnaisesti saadaan erilaisia etäisyyksiä, kuitenkin niin, että kohdesanaa lähempänä olevat sanat ovat useammin mukana kuin siitä kaukana sijaitsevat sanat. Intuitiivisesti voidaan ajatella, että niillä on todennäköisemmin enemmän merkitystä kuin kauempana sijaitsevilla sanoilla, joten niiden sisällyttäminen mukaan useammin käy järkeen.

Sanaston koko on rajattu, ja sitä merkitään v:llä. Yleensä sanaston koko on satoja tuhansia tai miljoonia. Tässä tutkielmassa on käytetty fastText-mallia, jossa sanaston koko on maksimissaan 22 500 000 sanaa. Tutkielmassa käytetyn sanaston koko oli 796 702 sanaa. Sanastoon valittavat sanat määräytyvät esiintymisten määrän perusteella, niin että mukaan otetaan v yleisintä sanaa, jotka kuitenkin esiintyvät aineistossa vähintään viisi kertaa. Toisin kuin monissa luonnollisen kielen käsittelyn sovelluksissa, yleisiä ja vähän merkitystä kantavia sanoja (stopwords) ei poisteta esikäsittelyvaiheessa, vaan myös ne sisällytetään sanastoon.

Opetukseen käytettävä neuroverkko on yksinkertainen. Siinä on vain kolme kerrosta: syötekerros, piilokerros ja ulostulokerros. Verkon opettaminen tapahtuu valetehtävän avulla. Se oppii joko arvaamaan sanan kontekstin perusteella (CBOW-malli) tai kontekstin sanan perusteella (skip-gram-malli). Tarkoituksena on saada kerrosten välisten kaarten painot sellaisiksi, että samaan kontekstiin liittyvät sanat ovat samalla tavalla painottuneita. Varsinaisesti kiinnostuneita ollaan kuitenkin vain syötekerroksen ja piilokerroksen välisten kaarten painokertoimista. Painot on esitetty matriisilla, jonka koko riippuu piilokerroksen koosta. Opetuksen jälkeen voidaan ottaa nämä painot ja hyödyntää niitä suoraan sanojen vektoriesityksinä, joiden pituus on piilokerroksen koko.

Syötekerroksessa sanat annetaan verkolle one-hot-encoding-muodossa. Skip-gram-mallissa syötteenä annetaan yksi sana, CBOW-mallissa syötesanojen määrä on $2c$. Piilokerroksen koko on d, missä d on yleensä joitakin satoja. Ulostulokerroksen koko on v eli sanaston määrä. CBOW-mallia opetettaessa katsotaan, onko ulostulo oppinut oikein yhden sanan, ja päivitetään mallia tämän oikean vastauksen perusteella. Skip-gram-mallissa tarkastellaan, onko ulostulo oppinut kaikkien sanaston sanojen osalta, kuuluvatko ne haluttuihin $2c$ sanaan. Sen jälkeen malleja voitaisiin päivittää jokaisen kontekstiin kuulumattoman sanan perusteella, mikä on kuitenkin käytännössä hidasta. Luvussa 2.2.2 kerrotaan tarkemmin, miten mallien opetus käytännössä tehdään. Mallien eroja on havainnollistettu kuvassa 1.

Skip-gram-malli ja CBOW-malli ovat siis varsin samanlaisia, sillä molemmissa sanoja edustaa konteksti, jossa ne esiintyvät. Käytännön kokeilla voidaan havaita, että usein toinen malli toimii hiukan paremmin tietyissä tehtävissä kuin toinen.



Kuva 1: CBOW ja skip-gram-mallien erot Mikolovia ym. [27] mukaillen. Kuva ei esitä verkon rakennetta, vaan havainnollistaa sitä, millaisilla syötteillä ja ulostuloilla mallit opetetaan.

2.2.1 Yksinkertaistettu versio mallista

Tarkastellaan aluksi yksinkertaistettua CBOW-mallia, jossa kontekstisanoja on vain yksi. Mallissa on siis yksi syötesana ja yksi ulostulosana. Myöhemmin nähdään, miten CBOW- ja skip-gram-mallit laajentavat tätä yksinkertaista mallia niin, että se toimii paremmin käytännössä. Mallia on havainnollistettu kuvassa 2.

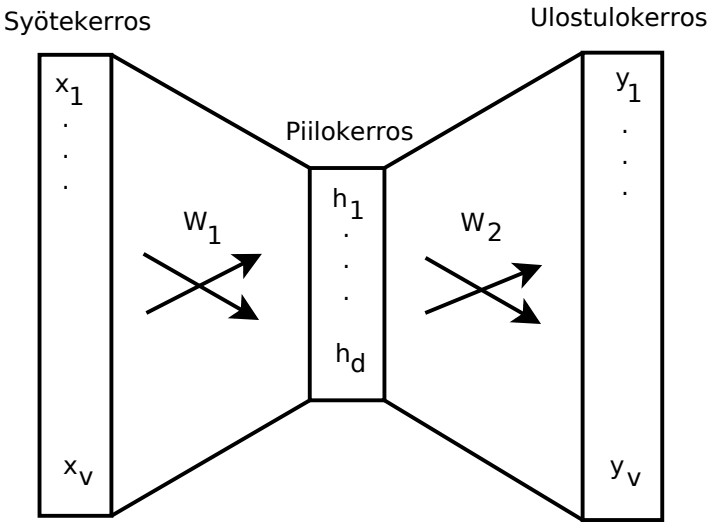
Yksinkertaisessa mallissa neuroverkko saa syötteenä yhden kontekstisanan w_c, joka esiintyy annetun sanan w_t ympärillä. Se palauttaa ulostulona kontekstiin kuuluvan kohdesanan. Syöte annetaan one-hot-encoding-muodossa.

Piilokerroksen koko on d. Syötekerroksen ja piilokerroksen väliset kaarten painot esitetään matriisilla W_1, jonka koko on $v \times d$. Matriisissa on siis d painoa jokaiselle sanaston sanalle, ja näitä painoja päivitetään verkon opetuksessa. Kun syöte luetaan, one-hot-vektori valitsee matriisista sanaa vastaavan rivin. Aktivaatiofunktiona piilokerros vain kopioi sanaa vastaavan vektorin, eli aktivaatiofunktio on identiteettifunktio $f(x) = x$.

Piilokerroksen ja ulostulokerroksen väliset painot esitetään toisella painomatriisilla W_2. Tämän matriisin koko on $d \times v$. Matriisista voi siis laskea jokaiselle sanaston sanalle todennäköisyyden sille, että se esiintyy annetussa kontekstissa. Ulostulokerroksen aktivaatiofunktio on softmax:

$$p(w_k|w_c) = \frac{e^{s(w_k,w_c)}}{\sum_{k'=1}^{v} e^{s(w_{k'},w_c)}}$$

missä s on jokin pisteytysfunktio, joista kerrotaan tarkemmin luvussa 2.3.1. Se tuottaa jokaiselle sanaston sanalle w_k aktivaation. Näitä aktivaatioita voidaan tulkita yhdessä todennäköisyysjakauman tavoin. Eli arvot ovat välillä



Kuva 2: Yksinkertainen malli, joka saa syötteenä yhden sanan ja jonka ulostulossa on myös vain yksi sana. Tässä v on sanaston koko, d on piilokerroksen koko, ja matriisit W_1 ja W_2 sisältävät kerrosten väliset painot.

$[0, 1]$ ja kaikkien ulostulokerroksen vektorin arvojen summa on yhteensä 1.

Neuroverkko tuottaa valetehtävään oikean tuloksen, jos sen mielestä oikea kohdesana on todennäköisin kohdesana annetulle kontekstisanalle, eli sanaa vastaavassa indeksissä on suurempi arvo kuin missään muussa indeksissä. Kun verkon tuottama tulos on selvillä, opetetaan verkkoa vastavirta-algoritmilla (backpropagation) eli optimoimalla kaarten painoja niin, että verkon tuottaman tuloksen etäisyys oikeasta tuloksesta pienenee [13].

Kun neuroverkko osaa suoriutua valetehtävästä hyvin, sen sivutuotteena tuottamat vektoriesitykset ovat myös "hyviä". Sanojen vektoriesitykset saadaan syötekerroksen ja piilokerroksen välisistä painoista eli matriisin W_1 riveistä. Jokaista sanaa vastaava vektoriesitys saadaan kertomalla matriisi sanan one-hot-esityksellä, ja vektorin pituus on d. Vaikka myös matriisissa W_2 on optimoituja painoja, siinä olevaa tietoa ei hyödynnetä.

2.2.2 CBOW ja skip-gram

CBOW- eli Continuous Bag of Words -malli opetetaan antamalla neuroverkolle tehtäväksi päätellä sana sen kontekstin perusteella. Toisin kuin yksinkertaisessa mallissa, kontekstisanoja ei ole vain yhtä, vaan $2c$. Eli kaikki enintään c:n etäisyydellä kohdesanasta ovat kontekstisanoja. Esimerkiksi kontekstista *kissa t maitoa* haluttaisiin päätellä, että puuttuva kohdesana t

on *juo*.

Koska syötesanoja on enemmän kuin yksi, identiteettifunktio ei kelpaa enää aktivaatiofunktioksi. Sen sijaan aktivaatiofunktio laskee kontekstisanojen one-hot-vektorien keskiarvon, eli se laskee vektorit yhteen ja jakaa saadun summan $2c$:llä. Muutoin CBOW-toimii luvussa 2.2.1 esitetyn yksinkertaistetun mallin mukaisesti.

Skip-gram-malli toimii tavallaan päinvastoin kuin CBOW-malli. Skip-gram-mallissa neuroverkko opetetaan päättelemään kontekstisanat kohdesanan perusteella. Tarkemmin sanottuna verkko kertoo jokaiselle sanaston sanalle todennäköisyyden sille, että sana on kohdesanan kontekstisana. Esimerkiksi sana *Korvatunturilla* esiintyy todennäköisemmin sanan *Joulupukki* tai sanan *lahja* kontekstissa kuin sanan *koala* tai sanan *kesäisin* kontekstissa.

Skip-gram-mallin alkupuoli toimii kuten yksinkertaistettu malli. Ulostulokerroksen kohdalla asia kuitenkin muuttuu, koska ei riitä tietää vain yhtä kontekstissa olevaa sanaa, vaan kontekstisanoja on $2c$. Skip-gram-mallin ulostulokerros on samanlainen matriisi kuin yksinkertaisessa mallissa, mutta sitä päivitetään kaikkien oikeiden vastausten osalta, ei vain yhden.

Oikeiden vastausten lisäksi skip-gram-mallissa ja CBOW-mallissa tulisi päivittää verkkoa myös väärien vastausten osalta, eli niiden sanaston sanojen, jotka eivät skip-gram-mallissa kuulu kohdesanan kontekstiin tai CBOW-mallissa ole kontekstin kohdesana. Käytännössä ei kuitenkaan ole järkevää päivittää mallia joka kerta jokaisen sanaston sanan perusteella. Sanaston koko on pienimmilläänkin kymmeniä tuhansia, joten verkon päivityksiä joutuisi tekemään hyvin paljon.

Yksi tapa ratkaista ongelma on käyttää *negatiivista näytteistystä* (negative sampling). Siinä ei päivitetäkään verkkoa jokaisen sanaston sanan perusteella, vaan valitaan satunnaisesti tietty määrä kielteisiä esimerkkisanoja eli sellaisia sanoja, jotka eivät ole oikeita vastauksia. Näin saadaan yksinkertaistettua tehtävä tavalliseksi luokittelutehtäväksi, jossa on positiivisia ja negatiivisia esimerkkejä.

2.3 fastText-malli

Skip-gram-malli ja CBOW-malli käsittelevät jokaista tekstissä esiintyvää sananmuotoa erikseen. Ne eivät ota huomioon sitä, että esimerkiksi *koiralla* ja *koirineen* ovat saman lekseemin kaksi eri sananmuotoa, vaan käsittelevät ne erillisinä. Tämä ominaisuus rajoittaa mallin laatua etenkin morfologisesti rikkaissa kielissä, kuten suomessa [4]. Lisäksi joissain tehtävissä datan riittävyys voi muodostua ongelmaksi: esimerkiksi puheentunnistuksen ja konekääntämisen alalla suuria määriä dataa ei usein ole saatavilla [26]. Yksi ratkaisu voisi olla hyödyntää morfologista tietoa vektoriesitysmallin opettamisessa, kuten turkin kielellä Sak ym. [35]. Tällainen malli vaatii kuitenkin usein kieliriippuvaista tietoa.

Bojanowskin ym. [4] *fastText-malli* on negatiivista näytteistystä hyödyntä-

vä skip-gram-malli, jota on laajennettu käyttämään sanojen vektoriesitysten muodostamisessa myös tietoa sanan alimerkkijonoista. Se siis vastaa morfologisen tiedon puutteen aiheuttamaan ongelmaan hyödyntämällä kohdesanan muodostavan merkkijonon alijonoja osana neuroverkon opetusta. Kohdesanan vektoriesitys muodostetaan merkkijonojoukon perusteella. Joukossa ovat sana itse sekä kaikki sen annetulla välillä olevat n-pituiset alimerkkijonot eli n-grammit.

Myös CBOW-kelpaa fastText-mallin pohjaksi. Tässä tutkielmassa esitetyt tulokset on kuitenkin saatu skip-gram-fastTextillä, joten se malli selostetaan tässä seuraavaksi. Skip-gram-malli valittiin siksi, että Bojanowskin ym. mukaan se toimii usein CBOW-mallia paremmin. Tutkielmaa varten tehtiin myös joitakin kokeita CBOW-fastTextillä, ja havaittiin skip-gram-mallin antavan parempia tuloksia.

2.3.1 Negatiivista näytteistystä käyttävä skip-gram-malli

Olkoon v sanaston koko. Jokainen sanaston sana on identifioitu indeksinsä mukaan, eli $w \in \{1, ..., v\}$. Tarkoituksena on oppia jokaiselle w:lle vektoriesitys. Jakaumahypoteesin perusteella halutaan oppia arvaamaan sana sen kontekstin perusteella. Eli formaalisti: otetaan opetusjoukoksi korpus, jossa on sanat $w_1, ..w_T$, ja skip-gram-mallin tarkoituksena on maksimoida seuraava logaritminen todennäköisyys:

$$\sum_{t=1}^{T} \sum_{c \in C_t} \log p(w_c | w_t),$$

missä konteksti C_t on niiden sanojen indeksien joukko, jotka ovat kohdesanan w_t ympärillä. Todennäköisyys sille, että kontekstisana w_c havaitaan kohdesanan w_t ympäristössä parametrisoidaan sanavektorien avulla. Ajatellaan nyt, että käytössä on pisteytysfunktio s joka kuvaa parin (sana, konteksti) reaaliluvulle.

Kontekstisanojen ennustamista voidaan ajatella joukkona erillisiä binäärisiä luokitteluongelmia. Tavoitteena on siis luokitella sanoja sen mukaan, ovatko ne kontekstissa vai eivät. Positiivisina esimerkkeinä toimivat kontekstissa esiintyvät sanat. Negatiivisina esimerkkeinä toimivat sanat arvotaan satunnaisesti sanastosta. Negatiivista näytteistystä käytettäessä ei siis oteta kaikkia sanaston muita sanoja negatiivisiksi esimerkeiksi, vaan nimenomaan satunnaisesti valittu osa.

Käytetään binääristä logistista häviöfunktiota, jolloin kontekstipositiolle c saadaan seuraava negatiivinen logaritminen todennäköisyys:

$$\log(1 + e^{-s(w_t, w_c)}) + \sum_{h \in H_{t,c}} \log(1 + e^{s(w_t, h)}),$$

missä $H_{t,c}$ on sanastosta poimittujen negatiivisten havaintojen joukko. Merkitään logistista häviöfunktiota seuraavasti: $f : x \mapsto \log(1+e^{-x})$. Tarkoituksena on minimoida häviö. Nyt tavoite voidaan kirjoittaa uudelleen muotoon:

$$\sum_{t=1}^{T}[\sum_{c\in C_t} f(s(w_t, w_c)) + \sum_{h\in H_{t,c}} f(-s(w_t, h))]$$

Tarvitaan siis enää pisteytysfunktio s. Luonnollinen tapa parametrisoida se on hyödyntää sanavektoreita. Olkoon jokaiselle sanaston sanalle w kaksi vektoria $u_w, v_w \in \mathbb{R}^d$. Näitä kahta vektoria kutsutaan joskus kirjallisuudessa syöte- ja ulostulovektoreiksi. Nyt voidaan sanoa, että meillä on kaksi vektoria u_{w_t} ja v_{w_c}, jotka vastaavat sanoja w_t ja w_c. Nyt pisteytys voidaan laskea skalaaritulona kohdesanan ja kontekstisanan vektoreina, siten, että $s(w_t, w_c) = u_{w_t}^\top v_{w_c}$. Pisteytys siis kertoo, ovatko sanavektorit samansuuntaisia. Näin tehtäisiin tavallisessa skip-gram-mallissa, jossa käytetään negatiivista valikointia. Kohta kuitenkin nähdään, että fastText-mallissa käytetään hiukan toisenlaista tulosfunktiota. Se hyödyntää tietoja alimerkkijonoista.

2.3.2 Alimerkkijonojen hyödyntäminen

Tässä tutkielmassa n-grammilla tarkoitetaan merkki-n-grammia, eli alimerkkijonoa. Sanan n-grammi on siis merkkijono, joka sisältyy yhtenäisenä sanaan ja on n merkkiä pitkä. Esimerkiksi sanan *silakka* 3-grammeja ovat muun muassa *ila*, *kka* ja *lak*. Käsitettä ei tule sekoittaa sana-n-grammiin eli n saneen peräkkäiseen esiintymiseen dokumentissa, vaikka n-grammilla voidaankin tarkoittaa sitä muissa yhteyksissä.

FastText-mallissa lisätään sanan alkuun $<$ ja loppuun $>$, jotta sanarajan erottaisi sanansisäisistä merkkijonoista. Mallissa sanan esitysjoukkoon voidaan valita eripituiset n-grammit joltakin väliltä. Siis valitaan kaikki alimerkkijonot, joiden pituus on välillä $[i, j]$ joillakin $i < j < l$, missä l on sanan pituus lisätyt aloitus- ja lopetusmerkit mukaan lukien. Jos siis sanassa on esimerkiksi 5 kirjainta, niin $l = 7$. Lisäksi mallin esitysjoukkoon otetaan mukaan koko sana, oli se minkä pituinen hyvänsä.

Esimerkiksi voitaisiin valita $i = 3$, $j = 5$ ja sanaksi *jänis*. Nyt n-grammijoukossa on merkkijonot {*<jä*, *jän*, *äni*, *nis*, *is>*, *<jän*, *jäni*, *änis*, *nis>*, *<jäni*, *jänis*, *änis>*, *<jänis>*}. Malli on siis hyvin yksinkertainen, eikä hyödynnä mitään tietoa siitä, mikä osa sanasta on sanavartalo tai taivutusmuoto. Se tekee mallista kieli- ja lisätiedosta riippumattoman.

Ajatellaan, että käytössä on n-grammien joukko G. Olkoon $G_w \subset G$ niiden n-grammien joukko, jotka esiintyvät sanassa w. Jokaisella $g \in G_w$ on vektoriesitys z_g. Sanat esitetään näiden vektoriesitysten summana. Saadaan siis pisteytysfunktio

$$s(w, c) = \sum_{g\in G_w} z_g^\top v_c \ .$$

Nyt sanojen esitykset muodostuvat osin niiden sisältämistä merkkijonoista, jolloin samoja merkkijonoja sisältävät sanat saavat samaa tietoa sisältävän esityksen. Siis tiedon jakaminen esitysten välillä on mahdollista, jolloin merkkijonoiltaan samankaltaiset sanat saavat samankaltaisemmat esitykset, ja malli tukee paremmin harvinaisten tai kokonaan opetusaineiston ulkopuolisten sanojen käsittelyä.

Alimerkkijonotietoa käytetään kuitenkin vain kohdesanan esityksen osana. Kontekstisanat esitetään tavalliseen tapaan erillisinä merkkijonoina.

2.3.3 Mallin toteutuksen yksityiskohtia

FastText-mallia opetettaessa käytetään kielteistä näytteistystä, eli kontekstissa esiintymättömistä sanaston sanoista valitaan satunnaisesti sanoja negatiivisiksi havainnoiksi. Bojanowski ym. [4] valitsivat negatiivisten esimerkkien määräksi 5. Konteksti-ikkunan koon c jakaumaksi valittiin tasajakauma väliltä $[1, 5]$. Ikkunan koko ei siis ole aina sama, vaan satunnaisuudesta johtuen se sisältää useammin sanoja, jotka ovat lähempänä kohdesanaa. Kun sanastoa luotiin, valittiin vain sellaiset saneet, jotka esiintyivät opetusaineistossa vähintään 5 kertaa. Sanaston koko voi olla maksimissaan 22 500 000 sanaa, ja ne on valittu yleisyyden perusteella. Tässä tutkielmassa sanaston koko oli 796 702 sanaa.

2.4 Mallin opettaminen

Tässä työssä on käytetty opetusaineistona suomenkielistä Wikipediaa esikuvina mm. Bojanowski ym. [4] ja Grave ym. [15]. Suomenkielinen XML-muotoinen Wikipedia-datasetti on ladattu 2.10.2017. Alkuperäisen suomenkielisen Wikipedia-datasetin koko on 2,4 gigatavua, mutta esiprosessoinnin jälkeen datasetistä jää jäljelle vain 882 megatavua, sillä suurin osa datasetistä on muuta kuin leipätekstiä. Lopulta fastText-kirjaston asetusten perusteella opetukseen käytettiin siitä vain 112 megatavua. 300-ulotteisen 3-6-grammeja hyödyntävän mallin opettaminen vei 118 minuuttia, kun opetukseen käytettiin Dell PowerEdge M610 -korttipalvelinta, jossa on 32 gigatavua keskusmuistia ja kaksi neliytimistä Intel Xeon E5540 -suoritinta. Opetusaika näyttäisi kasvavan lineaarisesti dimensionaalisuuden mukaan: samoilla n-grammeilla esimerkiksi 500-ulotteisen mallin opetus kesti 202 ja 700-ulotteisen 237 minuuttia.

2.4.1 Esiprosessointi

Bojanowski ym. [4] käyttivät Matt Mahoneyn englannin kielelle kehittämää esiprosessointiskriptiä. Skripti poistaa alkuperäisestä Wikipedia-datasetistä XML-merkinnät ja jättää jäljelle vain artikkelien otsikot ja tekstit.

Skripti korvaa numerot sanoilla (esimerkiksi *1* korvataan sanalla *one*), muuttaa isot kirjaimet pieniksi ja poistaa lopuksi kaikki merkit, jotka eivät

ole pieniä kirjaimia välillä *a* - *z*. Seurauksena aineisto on melko yhdenmukaista eikä merkistöongelmia synny, ja silti juurikaan informaatiota ei menetetä. Tosin vierassanoista poistetaan merkkejä, jotka eivät ole välillä *a* - *z*, ja moninumeroiset luvut kuvataan oikeasta lukutavasta poikkeaville sanoille (esimerkiksi *1991* kirjoitetaan muotoon *one nine nine one*, ei *nineteenhundred ninety nine* tai *thousand ninehundred ninety nine*.

Suomeksi ei kuitenkaan ole kannattavaa poistaa kaikkia merkkejä, jotka eivät ole välillä *a* - *z*, koska *ä* ja *ö* ovat suomen kielessä yleisiä ja merkitystä kantavia kirjaimia. Suomen kielessä myös numeraalit taipuvat, joten numeroiden korvaaminen lukusanoilla olisi vienyt niitä kauemmaksi oikeasta lukutavasta kuin englannin kohdalla. Koska suomen kielessä merkistöongelmat joutuu kohtaamaan joka tapauksessa, jätettiin tässä tutkielmassa myös muut aksenttimerkit kuin *ä*- ja *ö*-kirjaimissa esiintyvät paikoilleen. Suomen kielellä esiprosessoinnissa muutettiin siis isot kirjaimet pieniksi ja poistettiin ne merkit, jotka eivät ole numeroita tai kirjaimia.

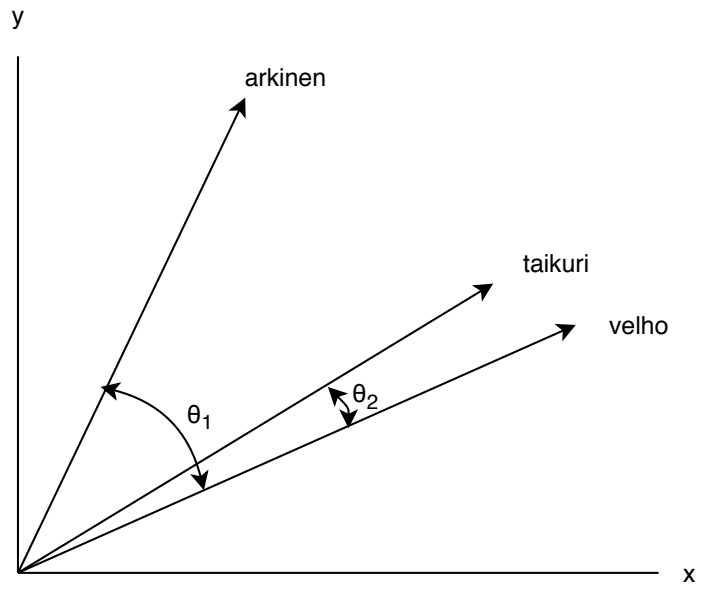
3 Semanttinen samankaltaisuus

Yksi tapa mitata fastTextillä luotujen mallien laatua verrattuna muihin malleihin on tutkia sen kykyä arvioida sanojen semanttista samankaltaisuutta [4]. Mallin voidaan sanoa toimivan hyvin, jos se kuvaa merkitykseltään samankaltaiset tai samoihin aihepiireihin kuuluvat sanat lähelle toisiaan ja merkitykseltään erilaiset ja eri aihepiireihin kuuluvat sanat kauas toisistaan. Mallin toimintaa arvioidaan vertaamalla sen muodostamien vektoriesitysten kosinietäisyyksiä ihmisten tekemiin arvioihin sanojen välisistä semanttisista etäisyyksistä. Samankaltaisuutta on havainnollistettu kuvalla 3.

3.1 Datasetit

Bojanowski ym. [4] testaavat fastTextiä semanttista samankaltaisuutta arvioivissa tehtävissä useilla eri kielillä. Datasettejä sanojen samankaltaisuuden ihmisarvioista on kuitenkin saatavilla vain vähän [19]. Tässä tutkielmassa samankaltaisuusarvioita on tehty kolmella eri datasetillä, joista yhdestä on käytetty kahta eri versiota. Käytetyt datasetit WS353 [10], siitä tutkielmaa varten lyhennetty WS277, SimLex300 [42] sekä FinnSim [42] esitellään tässä luvussa.

Samankaltaisuutta tutkiessa tarvitaan tietoa siitä, mitkä sanat ihmisarvioijat kokevat samankaltaisiksi. Tätä voidaan mitata näyttämällä arvioijille sanapareja ja kysymällä näiltä numeerista arvioita parin sanojen samankaltaisuudesta [34]. Samankaltaisuusarvioita tehtäessä sanat eivät esiinny muussa kontekstissa kuin kyselyn yhteydessä, eli sanoja ei näytetä esimerkiksi osana lausetta. Kyselyt näytetään kaikille arvioijille samassa järjestyksessä. Usein arviot teetetään kieltä äidinkielen tasoisesti osaavilla maallikoilla, mikä on



Kuva 3: Sanojen samankaltaisuus sana-avaruudessa. Sana *taikuri* on samankaltaisempi sanan *velho* kuin sanan *arkinen* kanssa, joten niiden välinen kulma θ_2 on pienempi kuin θ_1.

asiantuntija-arvioihin verrattuna yhtä toimiva menetelmä, kunhan arvioijia on tarpeeksi monta [38].

Kaikissa tässä tutkielmassa käytetyissä dataseteissä on hyödynnetty luokitusta asteikolla 0 - 10 korrelaatioita laskettaessa. Kuitenkin SimLex300-datasetin pohjana olevaa SimLex-999-datasettiä [18] koottaessa arvioinnit on tehty ohjeistuksella arvioida sanaparin samankaltaisuutta asteikolla 0 - 6. Jälkikäteen asteikko on laajennettu projisoimalla se välille 0 - 10, jotta datasetti vertautuisi paremmin muihin samankaltaisuusarvioissa sovellettaviin datasetteihin.

Ihmisten antamista arvioista on laskettu keskiarvo. Tätä keskiarvoa pidetään mittarina sanojen samankaltaisuudesta ihmisarvioijien mielestä. Korkea samankaltaisuusarvo on sanaparilla *tiikeri, tiikeri* (10,0), missä sanat ovat samat, mutta samankaltaisina pidettiin myös pareja *keskipäivä, puolipäivä* (9,29) ja *taikuri, velho* (9,0). Matala samankaltaisuusarvo on puolestaan esimerkiksi pareilla *kuningas, kaali* (0,23) ja *metsä, hautausmaa* (1,85). Jos samankaltaisuusarvo on 5,0 tai suurempi, voidaan sanoja pitää enemmän samankaltaisina kuin erilaisina.

Tässä tutkielmassa on käytetty samankaltaisuuden tutkimiseen Spearmanin järjestyskorrelaatiokerrointa [39], missä korrelaatio on laskettu ihmisarvioiden ja mallin antamien arvioiden välille. Spearmanin järjestyskorrelaatiokertoimella ainoastaan järjestyksellä on väliä, joten tässä tutkielmassa tarkoilla samankaltaisuusarvoilla ei ole väliä, kunhan vain tiedetään, mitkä

sanaparit ovat samanlaisempia kuin toiset. Jos korrelaatio on suuri, voidaan sanoa mallin toimivan hyvin.

Tutkielmassa haluttiin tarkastella suomenkielisen mallin toimimista verrattuna aiemman tutkimuksen muunkielisiin malleihin, joten WS353-datasetin kääntäminen oli tarpeellista.[1] Samankaltaisuusdatasettien kääntäminen ei ole poikkeuksellista, vaikka sitä vastaan on argumentoitu [23].

Jourbarne ym. [19] teettivät ranskaksi ja saksaksi käännetyllä datasetillä uuden ihmisarvion kohdekielen puhujilla. Vastaavasti toimi saksaksi käännetyllä datasetillä Gurevych [16] ja Panchenko ym. [29] venäjänkielisellä käännöksellä WS353-datasetistä. Myös uudelleenarvioimattomia käännöksiä englanninkielisistä dataseteistä on käytetty, ja ne voivat toimia alkukielisiin vertautuvalla tavalla, kuten Postman ym. hollanninkielinen käännös [31].

Useimmissa käännöksissä on törmätty käännösongelmiin, jotka ovat johtaneet painottuneiden valintojen tekemiseen tai hankalien sanojen poistamiseen käännetyistä dataseteistä. Myös yksikäsitteistä kääntämistä monimutkaisempia menetelmiä on käytetty, kuten Agirre ym. [1], jotka hyödynsivät espanjankielisten sanavektorien tutkimuksessa konekääntämistä ja ottivat huomioon useita eri käännösvastineita tutkittujen sanavektorien tarkkailussa yhden käännösvastineen vektorin sijaan.

3.1.1 WS353 ja WS277

Tunnettu englanninkielisten mallien toimivuuden tutkimiseen käytetty datasetti on yleiskielisistä sanoista koostuva WS353 [10]. Tässä tutkielmassa on käytetty tutkielmaa varten suomennettua versiota. WS353 on kerätty kahdessa osassa. Ensimmäinen puolisko datasetistä sisältää 153 sanaparia sekä 13 ihmisen arvion niiden sanojen samankaltaisuudesta. Toinen puolisko datasetistä sisältää 200 sanaparia, joita arvioimassa oli 16 ihmistä. Arvioinnit annettiin asteikolla 0 - 10. Pistetys 0 merkitsee, että sanat eivät lainkaan liity toisiinsa (“totally unrelated words”). Pisteytys 10 tarkoittaa, että sanat liittyvät toisiinsa hyvin kiinteästi tai että ne ovat identtiset (“very much related or identical words”).

Yhteensä datasetissä on 353 sanaparia. Suurin osa sanoista on substantiiveja ja yleisnimiä, kuten *kuvernööri*, *salaisuus* ja *keskiviikko*. Jotkin substantiiveista ovat erisnimiä, kuten *Mars* ja akronyymejä, kuten *FBI*. Yksi sanoista, *älykäs*, on adjektiivi. Jotkin sanat voivat olla englanniksi joko verbejä tai substantiiveja, mutta ne on käännetty substantiiveiksi. Esimerkiksi *love* on suomennettu *rakkaus*, ei *rakastaa*. Lähes kaikki aineiston sanat ovat yksikössä ja perusmuodossa. Englanninkielisessä datasetissä on kaksi sanaa, jotka eivät ole yksikössä, koska ne esiintyvät (halutussa merkityksessä) ainoastaan monikossa: *physics* ja *news*. Suomeksi ainoastaan *uutiset* on monikossa, sillä vaikka myös *uutinen* on suomenkielen sana, on kään-

[1] Käännetyt datasetit löytyvät osoitteesta https://github.com/adaisti/fin-eval.

nöksessä pyritty säilyttämään datasetti mahdollisimman samanlaisena kuin alkuperäinen.

Joitakin sanoja ei voi kääntää suomeksi yksikäsitteisesti, sillä niillä voi olla monta merkitystä, eivätkä ne testissä esiinny merkitystä implikoivassa lausekontekstissa. Esimerkiksi *stock*, joka voidaan kääntää suomeksi asiayhteydestä riippuen muun muassa *osake*, *varasto*, *varsi*, *kalusto*, *liemi* tai *karja*. Tällaisissa sanoissa on valittu sellainen käännösvastine, joka liittyy samaan aihepiiriin kuin sanaparissa esiintyvä toinen sana. Esimerkiksi *stock* on käännetty sanan *market* (*markkinat*) yhteydessä *osake*. Käännösratkaisussa sanaparin toinen sana on siis nähty käännettävän sanan kontekstina, sillä myös ihmisarvioija on nähnyt sanat samalla kertaa. Kontekstin näkemisestä kertoo myös englanninkielisten arvioijien antamat korkeat samankaltaisuusarviot: parin *stock*, *market* samankaltaisuusarvioiden keskiarvo on 8,0.

Testitilanteessa sanaparit on näytetty arvioijille aina samassa järjestyksessä, joten myös edeltävät sanaparit vaikuttavat koehenkilöiden käsityksiin sanojen merkityksestä. Tämän huomioiden esimerkiksi *stock, market*-sanaparia seuraavalla rivillä olevaan sanapariin *stock, CD* valittiin myös sanan *stock* käännösvastineeksi *osake*. Sanalle *stock* ei muodostu muuta merkitystä tukevaa kontekstia sanasta *CD*, eivätkä ihmisarvioijat ole pitäneet sanoja kovin samankaltaisina: samankaltaisuusarvioiden keskiarvo on 1,31.

WS535-datasetti sisältää myös sanoja, jotka ovat englanniksi synonyymejä ja joilla on vain yksi suomenkielinen käännösvastine. Esimerkiksi *rooster* ja *cock* kääntyvät molemmat suomeksi *kukko*. Kuten Granada ym. [14] ja Joubarne ym. [19], tällaisille sanoille on käytetty samaa käännösvastinetta, vaikka lähdekielen sanat olisivatkin olleet eri. Tämän takia datasetistä löytyy sana *kukko* sekä parista *ruoka, kukko* (alun perin *food, rooster*) että *lintu, kukko* (alun perin *bird, cock*), vaikka alun perin pareissa ei ole yhtään samaa sanaa. Tässäkin käännöksessä on hyödynnetty tietoa samankaltaisuusarviosta: vaikka englannin kielen sanalla *cock* on muitakin merkityksiä (kuten *hana*), kertoo samankaltaisuusarvioiden keskiarvo 7,10 arvioijien yhdistäneen sanan voimakkaasti *lintuun*.

Joillekin WS353-datasetin sanoille ei ollut mahdollista löytää merkityskentältään tarpeeksi samankaltaista käännösvastinetta. Monet WS353-datasetin sanat on tarkoituksella valittu sellaisiksi, että niissä esiintyy polysemiaa eli samasta lähteestä olevan sanan erilaisia semanttisia merkityksiä. Samoin datasetissä esiintyi sanaliittoja eli kahdesta sanasta muodostuvia fraaseja.

Esimerkiksi sanaparilla *bird, crane* jälkimmäinen sana voi yhdistyä kahteen käännösvastineeseen *kurki* ja *nostokurki*. Sanan käännösvastineeksi valitsin *kurki*, koska epäilin sen tulleen arvioijille useammin mieleen, sillä sanojen samankaltaisuus oli suuri (keskimäärin 7,38). Kääntämisen vaikeus korostui tilanteissa, joissa sanaparin sanat oli koettu erilaisiksi, eikä toinen sana tarjonnut kontekstia kääntämiselle. Esimerkki tästä tilanteesta on pari *stock, jaguar*, sillä on mahdotonta sanoa, miten arvioijat tulkitsivat sanan *stock*, joka voi tarkoittaa esimerkiksi *osake*, *varasto*, *karja* tai *pölkky*. Kään-

nöksessä valittiin merkitys *osake*, sillä sana esiintyi myös parin *stock, market* osana ja sama käännös haluttiin säilyttää kaikissa datasetin sanapareissa.

Jotkin WS353-datasetin sanat ovat sellaisia synonyymejä eli lähes samaa tarkoittavia sanoja, ettei niille löydy suomeksi kuin yksi luonteva käännösvastine. Tämän takia esimerkiksi sanapari *gem, jewel* käännettiin vähemmän yleisillä termeillä *korukivi, jalokivi*, vaikka *jalokivi* olisi prototyyppisempi käännösvastine kummallekin sanaparin sanalle. Toinen esimerkki synonymiasta on *coast, shore*, joka käännettiin *rannikko, ranta*.

Joissakin sanapareissa jouduttiin kääntäessä ottamaan kantaa sanaluokkaan. Esimerkiksi sanaparissa *tool, implement* sana *implement* voitaisiin kääntää *toteuttaa*, mutta käännökseksi valittiin *väline*, sillä on vaikea esittää arviota kahden eri sanaluokan sanan välisestä samankaltaisuudesta. Todennäköisesti arvioijat siis kokivat sanat myös molemmat substantiiveina.

Jotkin sanaparit muodostivat englanniksi sanaliiton, eli niistä muodostuu jokin käsite. Esimerkiksi sanaparille *death, row* voi muodostua tulkinta *death row*, suomeksi *kuolemanselli*. Kuitenkin erillään sanaparin sanat merkitsevät *kuolema, rivi*, eli sanat eivät juurikaan liity samaan kontekstiin. Toinen esimerkki vastaavasta sanaparista on *soap, opera*, joka saa sanaliittona tulkittuna merkityksen *soap opera*, *saippuaooppera*, vaikka yksittäiset sanat eivät juuri liity toisiinsa. Kuitenkin arvioijat kokivat sanaparit enemmän samankaltaisina kuin erilaisina: *death, row* sai samankaltaisuusarvioksi 5,25 ja *soap, opera* jopa 7,94.

Jotkin sanaparit eivät olleet itsessään ongelmallisia, mutta niissä jouduttiin käyttämään samaa käännösvastinetta kuin toisen sanaparin eri sanalle. Esimerkiksi sanapari *game, victory* oli ongelmallinen, koska *victory* kääntyy *voitto*, kuten myös sanaparin *profit, loss* ensimmäinen sana. Vaarana on, että suomenkielinen arvioija olisi samojen sanojen perusteella kenties tehnyt erilaisen arvion sanoista aiemmin nähtyjen kontekstien perusteella. Joissakin peräkkäisissä kyselyissä arvioijille näytettiin samaa sanaa eri sanaparien osana, joten aiemmin nähtyjen sanojen konteksti on vaikuttanut englanninkielisten arvioijien antamiin arvioihin.

WS353 on siis kattava käännös alkuperäisestä englanninkielisestä datasetistä, mutta suomen kielen kannalta se on osittain ongelmallinen ylempänä esitettyjen ongelmatapausten vuoksi. WS277 on lyhennetty versio samasta datasetistä, ja siinä ongelmalliset sanaparit on poistettu. Se ei siis sisällä yhtään tässä luvussa esitetyistä ongelmatapauksista, vaan sen sanoille löydettiin riittävän yksikäsitteinen ja prototyyppinen käännös.

3.1.2 SimLex

Toinen alun perin englannin kielelle kehitetty ja myöhemmin osittain suomeksi käännetty datasetti on Hill ym. esittelemä SimLex-999-datasetti [18]. Datasetti sisältää 999 sanaparia. SimLex-999-datasettiä luodessa ihmisarvioijia ohjeistettiin pisteyttämään sanaparit sen mukaan, ovatko ne samankaltaisia

(“similar”). Sanaparit eivät siis saaneet korkeaa pistemäärää, jos niiden sanat liittyivät samaan konseptiin (“relatedness”) mutta eivät tarkoittaneet samaa asiaa. Esimerkiksi sanapari *clothes, closet* liittyy samaan konseptiin, mutta sen sanojen merkitys on hyvin erilainen [17]. SimLex-999-datasetissä sen luokitus on 1,96, kun WS353-datasetissä saman sanaparin luokitus on 8,00. Sen sijaan samaa merkitsevät sanaparin *coast, shore* sanat, jonka luokitus on SimLex-999:ssä 9,00 ja WS353:ssa 9,10. Tässä SimLex-999:n luokitukset on projisoitu asteikolle 0 - 10, vaikka ne alun perin kerättiin asteikolla 0 - 6.

Tässä tutkielmassa on käytetty 300 sanaparin pituiseksi lyhennettyä ja suomeksi käännettyä SimLex-300-datasettiä [42] (jäljempänä SimLex), jossa on 300 sanaparia. Senkin sisältämät arviot ovat englanninkielisellä datasetillä tehtyjä niin kuin WS353- ja WS277-dataseteissä. Koko SimLex-999-datasettiä ei siis ole käytetty tutkielman aineistona.

3.1.3 FinnSim

Suomenkielisten arvioijien tekemiä arvioita sisältää Venekosken ym. [42] esittelemä FinnSim300-datasetti (jäljempänä FinnSim). Datasetin sanaparit perustuvat Hill ym. esittelemään SimLex-999-datasettiin [17]. FinnSimiin on valittu 300 sanaparia SimLexistä, joiden samankaltaisuutta arvioi 55 suomea äidinkielenään puhuvaa arvioijaa. siis FinnSim on ainoa tutkielmassa käytetty suomenkielisten arvioijien avulla muodostettu datasetti. SimLexin mukaisesti arvioijia ohjeistettiin antamaan täydet pisteet, mikäli sanat ovat synonyymeja, ja 0, mikäli sanat ovat täysin erilaiset. Kuten SimLex-999-datasetin kohdalla, merkitystä oli sanojen samankaltaisuudella, ei samaan konseptiin liittymisellä. Sanojen vektoriesitysmallit ovat aiemmassa tutkimuksessa korreloineet huonommin SimLex-datasetin sanaparien kuin WS353:n sanaparien kanssa, eli tehtävä on vektoriesitysmallille vaikeampi kuin samaan konseptiin liittyminen [17]. Kuten myöhemmin esitetään, tässä tutkielmassa havaittiin sama ilmiö sekä SimLex- että FinnSim-datasetin kanssa.

3.2 Tulokset

Tätä tutkielmaa varten testattiin semanttista samankaltaisuutta eri dataseteillä, erikokoisilla sanavektorien ulottuvuuksilla d ja eri n-grammien pituusväleillä. Kuten Bojanowski ym. [4] tarkoituksena oli selvittää, millaisilla parametreilla malli toimii parhaiten, ja parantaako alimerkkijonojen käyttö sen laatua.

Kokeissa muodostettiin kielimalleja fastText-kirjaston[2] avulla eri parametreilla. Parametreja pyrittiin kokeilemaan mahdollisimman laajasti. Tämän takia sekä vektorin ulottuvuuksien että n-grammin pituuksien kohdalla kokeiltiin myös suurempia arvoja kuin Bojanowski ym. kokeissa. Osa hyvistä tuloksista sijoittuikin suurempien arvojen alueelle.

[2] https://github.com/facebookresearch/fastText

Taulukko 1: Datasettien Spearmanin korrelaatiokerroin erilaisilla malleilla. d on sanavektorin ulottuvuuksien määrä ja n-grammi on mallin käyttämä alimerkkijonotieto: 0-grammi tarkoittaa, että alimerkkijonoja ei käytetty. Taulukossa on lihavoitu paras yhdistelmä.

(a) WS353.

	n-grammi		
d	0	3 - 6	4 - 7
100	0,523	0,553	0,585
200	0,540	0,585	0,599
300	0,545	0,579	0,612
400	0,558	0,594	0,605
500	0,536	0,575	**0,617**
600	0,536	0,589	0,614
700	0,538	0,588	0,608

(b) WS277.

	n-grammi		
d	0	3 - 6	4 - 7
100	0,591	0,637	0,659
200	0,601	0,666	0,679
300	0,617	0,665	**0,688**
400	0,624	0,671	0,681
500	0,614	**0,688**	**0,688**
600	0,605	0,662	0,671
700	0,605	0,664	0,670

(c) SimLex.

	n-grammi		
d	0	3 - 6	4 - 7
100	0,210	0,210	0,232
200	0,235	0,232	0,241
300	0,253	0,249	0,252
400	0,242	0,244	0,272
500	0,245	0,260	0,291
600	0,243	0,274	**0,297**
700	0,245	0,263	0,287

(d) FinnSim.

	n-grammi		
d	0	3 - 6	4 - 7
100	0,183	0,191	0,214
200	0,200	0,216	0,225
300	0,218	0,232	0,243
400	0,224	0,239	0,245
500	0,228	0,247	**0,287**
600	0,217	0,254	0,279
700	0,217	0,245	0,275

Samankaltaisuutta mitattiin neljällä datasetillä: samaan kontekstiin kuuluvien sanojen datasetillä WS353 ja sen osajoukolla WS277 sekä samaa tarkoittavien sanojen dataseteillä SimLex ja FinnSim. Tulokset on esitetty taulukoissa 1 ja 2. Seuraavissa alaluvuissa käydään tulokset läpi yksityiskohtaisesti.

3.2.1 Vektorin ulottuvuuksien määrä

Vektoriesitysmallia muodostettaessa pitää valita, kuinka moniulotteisia vektoreita on tarkoitus tuottaa. Tuotettavien vektorien koko on verkon piilokerroksen koko. Kuten aiemmin ulottuvuudesta käytetään merkintää d. Jos ulottuvuusmäärä on pieni, mallin esitysvoima jää myös pieneksi eikä se pysty välttämättä pysty ilmaisemaan semanttisia eroja. Kuitenkin suurempiulotteisten vektoriesitysten opettaminen vie kauemmin aikaa [26]. Merkitään i:llä minimi-n-grammia ja j:llä maksimi-n-grammia ja i-j-grammilla n-grammeja i:stä j:hin. Siis jos esim. $i = 3$ ja $j = 5$, niin mukana 3-5-grammeissa ovat

3-grammit, 4-grammit ja 5-grammit.

Tutkielmaa varten tuotettiin vektoreita, joilla $d \in \{100, 200, ..., 700\}$. Ulottuvuuden vaikutusta testattiin sekä tavalla, jossa ei hyödynnetty alimerkkijonoja, että kahdella tavalla n-grammeja hyödyntäen: 3-5-grammeilla sekä 4-7-grammeilla. Tulokset on esitetty taulukossa 1.

N-grammien hyödyntämisestä riippumatta ulottuvuuden kasvattaminen 100-kokoisissa askelissa lisäsi aluksi mallin laatua, mutta sitten laski sitä. Ulottuvuutta kasvattaessa löytyy siis huippukohta, jonka jälkeen useammat ulottuvuudet huonontavat laatua. Näiden datasettien kohdalla huippukohta on noin 500.

Mallin ulottuvuuksista riippumatta alimerkkijonojen hyödyntäminen paransi melkein aina mallin laatua. Kokeissa ei löytynyt yhtään ulottuvuutta, jolla alimerkkijonoja hyödyntämätön malli olisi aina pärjännyt paremmin kuin 3-6-grammeja tai 4-7-grammeja hyödyntävä malli. Kuitenkin SimLex-datasetillä ja pienillä ulottuvuusmäärillä alimerkkijonoja hyödyntämätön malli pärjäsi hyvin. Kuten taulukosta 1c nähdään, 300-ulotteisessa tapauksessa sen korrelaatiokerroin on 0,253 eli yhtä hyvä kuin 4-7-mallin, kun 3-6-grammi-mallin on 0,249.

Havaitaan, että ulottuvuuksien kasvaessa n-grammi-mallit saivat kuitenkin selvästi paremmat tulokset. Ulottuvuuksien määrä on siis yhteydessä n-grammien hyödyntämisen kannattavuuteen. Esimerkiksi 600 ulottuvuudella alimerkkijonoja hyödyntämätön malli saa SimLex-datasetillä korrelaatiokertoimeksi 0,243, 3-6-grammi-malli 0,274 ja 4-7-grammi malli 0,297.

3-6-grammeja hyödyntävä malli oli myös lähes kaikissa kokeissa 4-7-grammeja hyödyntävää mallia huonompi: ainoastaan WS277-datasetillä (taulukko 1b) ja 500-ulotteisilla vektoreilla mallit saivat saman tuloksen, 0,688, joka on myös paras korrelaatiokerroin kokeilluilla parametreilla.

Kokeiden perusteella hyvä ulottuvuus vektoreille on 500, koska sillä mallin laatu on varsin hyvä datasetistä riippumatta. WS353- ja WS277-dataseteillä myös 300-ulotteiset vektorit toimivat lähes yhtä hyvin kuin 500-ulotteiset. Kokeiden perusteella voidaan sanoa, että alimerkkijonojen käyttäminen parantaa mallia vektorien ulottuvuuksista riippumatta.

3.2.2 N-grammien pituus

Tässä tutkielmassa haluttiin vertailla, millä välillä n-grammien tulisi olla, jotta malli toimisi mahdollisimman hyvin. Tätä varten kokeiltiin eri väleillä olevia n-grammien arvoja. Kuten edellä, käytetään merkintää i minimi-n-grammista ja j maksimi-n-grammista, ja i-j-grammilla merkitääm n-grammeja niiden väliltä.

Bojanowski ym. [4] valitsivat useimpien kokeidensa parametreiksi $i = 3$ ja $j = 6$ ja testasivat erilaisia i:n ja j:n arvoja välillä $[2, 6]$ saksan- ja englanninkielisillä dataseteillä. Näillä dataseteillä 3-6-grammit toimivat melko hyvin, mutta eri kieliä varten eri parametrit voivat olla paremmat.

Taulukko 2: Datasettien Spearmanin korrelaatiokerroin eri n-grammiväleillä, $d = 300$, i on minimi ja j maksimi.

(a) WS353.

$i\backslash j$	2	3	4	5	6	7	8	9	10
2	**0,473**	0,471	0,542	0,548	0,554	0,580	0,572	0,560	0,569
3		**0,491**	0,553	0,584	0,579	0,585	0,576	0,578	**0,594**
4			**0,590**	0,594	**0,604**	**0,612**	0,601	0,597	0,610
5				**0,598**	0,587	0,596	0,598	**0,617**	0,587
6					0,590	0,599	**0,600**	0,597	0,599
7						0,592	0,578	0,587	0,582
8							0,585	0,579	0,583
9								0,575	0,564
10									0,575

(b) WS277.

$i\backslash j$	2	3	4	5	6	7	8	9	10
2	**0,537**	0,542	0,633	0,637	0,637	0,663	0,659	0,639	0,652
3		**0,576**	0,636	0,669	0,665	0,665	0,655	0,655	**0,672**
4			**0,673**	**0,676**	**0,684**	**0,688**	**0,678**	**0,675**	0,686
5				0,670	0,654	0,664	0,670	0,673	0,663
6					0,661	0,661	0,653	0,663	0,665
7						0,645	0,643	0,661	0,651
8							0,655	0,644	0,647
9								0,650	0,636
10									0,653

(c) SimLex.

$i\backslash j$	2	3	4	5	6	7	8	9	10
2	**0,168**	0,193	0,200	0,227	0,229	0,229	0,226	0,212	0,225
3		**0,222**	0,221	0,222	0,249	**0,267**	0,247	0,251	0,246
4			**0,240**	**0,247**	0,251	0,252	0,251	**0,261**	0,255
5				0,246	0,245	0,245	**0,263**	0,257	0,259
6					**0,271**	0,257	0,259	0,256	**0,262**
7						0,252	0,251	0,251	0,244
8							0,258	0,251	0,250
9								0,244	0,225
10									0,241

(d) FinnSim.

$i\backslash j$	2	3	4	5	6	7	8	9	10
2	**0,152**	0,179	0,202	0,218	0,208	0,209	0,206	0,197	0,211
3		**0,218**	0,224	0,224	0,232	**0,250**	**0,231**	0,233	0,232
4			**0,245**	0,232	0,233	0,243	0,251	**0,259**	0,232
5				**0,236**	0,227	0,233	0,242	0,239	**0,245**
6					**0,246**	0,229	0,231	0,233	0,233
7						0,215	0,216	0,218	0,216
8							0,217	0,213	0,224
9								0,206	0,207
10									0,209

Tässä tutkielmassa arvoja on kokeiltu välillä $[2, 10]$. Yhtenä syynä on hypoteesi siitä, että pidemmätkin n-grammit kantavat enemmän merkitystä suomen kielessä, koska suomen sananmuodot ovat keskimäärin pidempiä kuin englannin. Toinen syy on se, että kokeita tehdessä havaittiin n-grammien ja erityisesti j:n kasvattamisen lisäävän vektoriesitysten etäisyyksien korrelaatiota ihmisarvioijien antamaan samankaltaisuusarvioihin. Kaikissa kokeissa $d = 300$, sillä se antoi kohtalaisen hyviä tuloksia edellisen luvun ulottuvuuskokeissa kaikilla dataseteillä ja vie vähemmän opetusaikaa kuin hiukan paremmin toiminut $d = 500$. Lisäksi myös Bojanowskin ym. [4] kokeissa käytettiin $d = 300$, joten valinta helpottaa tulosten vertailua aiempaan tutkimukseen.

Tulokset on esitelty taulukossa 2. Taulukon rivit kertovat i:n arvon ja sarakkeet j:n arvon. Taulukoissa on korostettu yksi lukuarvo jokaisella rivillä sen mukaan, mikä tuloksista on paras. Eli jokaista i:n pituutta kohti on korostettu, minkä pituinen j antaa parhaan tuloksen.

Kuten taulukoista käy ilmi, pitkien n-grammien sisällyttäminen mukaan esitykseen parantaa esityksen laatua. Erityisesti taulukossa 2a nähdään, että WS353-datasetillä i:stä riippumatta j:n pitäisi olla vähintään 7, jotta saadaan paras mahdollinen lopputulos. Tulokset eivät kuitenkaan kasva lineaarisesti, kun j:n arvoa kasvatetaan. Esimerkiksi 5-9-grammit sisällyttämällä korrelaatio on 0,617, mutta kun 10-grammit otetaan mukaan, korrelaatio putoaa arvoon 0,587. Osittain tämä johtuu todennäköisesti myös satunnaisvaihtelusta.

Myös muilla dataseteillä havaitaan, että esitys hyötyy usein suurista j:n arvoista. Etenkin, kun pienet i jätetään pois, saadaan paras korrelaatio ottamalla mukaan myös 10-grammit. Esimerkiksi WS277-datasetillä (taulukko 2b), kun $i = 6$, paras tulos saadaan valitsemalla j:n arvoksi 10.

Pienien n-grammien ottaminen mukaan huonontaa usein tulosta verrattuna niiden jättämiseen pois. Pääsääntöisesti $i = 2$ ja $i = 3$ antavat huonompia tuloksia kuin $i = 4$: ainoastaan FinnSim- ja SimLex-datasetillä (taulukot 2d ja 2c) havaintaan, että kun $j = 7$, $i = 3$ antaakin paremman tuloksen kuin $i = 4$.

WS353-datasetillä (taulukko 2a) parhaan tuloksen saa valitsemalla esitykseen mukaan 5-9-grammit, jolloin korrelaatio on 0,617. Lähes yhtä hyvin toimivat 4-7-grammit korrelaatiolla 0,612. WS277-datasetillä (taulukko 2b), joka on WS353-datasetin osajoukko, 4-7-grammit antavat parhaan tuloksen (0,688), ja 5-9-grammit toiseksi parhaan (0,673).

FinnSim- ja SimLex-datasetit ovat eriluonteisia ja vaikeampia kuin WS353 ja WS277, mikä näkyy myös tuloksissa. Suhteelliset erot eri n-grammien valintojen välillä ovat ensiksi mainituilla suurempia. WS353-datasetillä pienin kokeissa löydetty korrelaatioarvo on 0,473 ja suurin 0,617. WS277-datasetillä nämä luvut ovat 0,537 ja 0,688, eli suhteellinen ero pienimmän ja suurimman arvon välillä on vieläkin pienempi. Sen sijaan FinnSim-datasetillä pienin löytynyt arvo on 0,152 ja suurin 0,259, SimLexillä 0,168 ja 0,271.

FinnSim-datasetillä (taulukko 2d) huomataan kaikista voimakkaimmin olleen hyötyä myös pienten n-grammien mukaan ottamisesta. Parhaat korrelaatiotulokset on saatu, kun $i = 4$: 4-9-grammeilla korrelaatio on 0,259 ja 4-8-grammeilla 0,251. Kolmanneksi paras tulos on saatu hyödyntämällä 3-7-grammeja. Myös SimLex-datasetin (taulukko 2c) kanssa 3-7-grammit toimivat hyvin: niillä korrelaatio on 0,267.

SimLex-tapauksessa kaikista paras korrelaatio on kuitenkin saatu 6-6-grammeilla, joka siis hyödyntää ainoastaan 6 merkkiä pitkiä n-grammeja sanan itsensä lisäksi. Näillä parametreilla korrelaatio on 0,271, kun ilman alimerkkijonoja korrelaatioksi saadaan 0,253 (taulukko 1c).

Korrelaatioita 0,267 ja 0,271 ei voida yleisesti ottaen pitää kovin korkeina, vaikka ne ovatkin tämän tutkielman piirissä parhaita saatuja tuloksia. Kuitenkin aiemmassa kirjallisuudessa esitetyt korrelaatiot samoille dataseteille ovat samaa luokkaa tai huonompia, kuten seuraavassa luvussa nähdään.

3.2.3 Tulosten vertailua

WS353-datasetillä sekä sen osajoukolla WS277-datasetillä huomataan, että alimerkkijonojen hyödyntäminen parantaa korrelaatiota enemmän verrattuna pelkästään vektorin ulottuvuuksien määrän optimoimiseen. WS353-datasetillä paras kokeissa havaittu korrelaatiotulos on 0,617. Tuloksen saatu kahdella tavalla: hyödyntämällä sekä useampiulotteisia vektoreita että sopivia n-grammeja ja pelkästään sopivien n-grammien avulla 300-ulotteisilla vektoreilla. Sopivat yhdistelmät ovat 500-ulotteinen vektori ja 4-7-grammit sekä 300-ulotteinen vektori ja 5-9-grammit. Myös 300-ulotteisilla vektoreilla ja 4-7-grammeilla päästään lähellä olevaan tulokseen 0,612.

WS277-datasetillä paras havaittu tulos on 0,688. Myös tähän tulokseen on päästy usealla eri tavalla: 300-ulotteisilla vektoreilla se saadaan valitsemalla 4-7-grammit, mutta myös 500-ulotteisilla vektoreilla se saadaan 4-7-grammien sekä 3-6-grammien avulla. 300-ulotteisilla vektoreilla saadaan lähellä olevat tulokset 0,686 4-10-grammeilla ja 0,684 4-6-grammeilla.

WS353- ja WS277-datasettien tulokset muistuttavat toisiaan, koska WS277-datasetti on muodostettu WS353-datasetin pohjalta. WS277-datasettiä koskevat tulokset ovat kauttaaltaan parempia kuin WS353-datasetin, koska sitä muodostaessa on jätetty suomen kielen kannalta ongelmalliset sanat pois. Sitä koskevat tulokset kuvaavat korrelaatiota ihmisten tekemien arvioiden kanssa paremmin kuin WS353-datasetin, ja tehtävä on siksi helpompi. Tämän vuoksi sitä koskevat tulokset vertautuvat paremmin englannin kieltä tai muita WS353-datasetin käännöksiä koskeviin tuloksiin, sillä se on hengeltään enemmän WS353-datasetin luonteinen.

Kummallakaan datasetillä ei kuitenkaan päästä vielä Bojanowski ym. [4] saavuttamiin englannin kieltä koskeviin tuloksiin. Bojanowski ym. saivat englannin kielellä datasetin WS353 korrelaatioksi alimerkkijonoja hyödyntämällä 0,71, kun vektorin pituus oli 300 ja n-grammeina käytettiin 3-6-

grammeja. Se on huomattavasti korkeampi kuin suomen kielen 0,579, joka on saatu samoilla parametreilla, tai 0,617, joka on saatu parhailla mahdollisilla parametreilla. Toisaalta Bojanowskin ym. kokeissa alimerkkijonojen hyödyntäminen ei tarjonnut parasta mahdollista tulosta englanninkielisellä mallilla, vaan alimerkkijonoista piittaamaton CBOW-malli sai korkeimman korrelaation (0,73) ja myös skip-gram-malli toimi paremmin ilman alimerkkijonoja (0,72). Suomen kielelle skip-gram-mallin vastaava arvo on 0,545, mikä on selvästi pienempi kuin 3-6-grammeja hyödyntävä 0,579.

Bojanowski ym. kokeilivat WS353-datasettiä myös muille kielille käännettynä. Eniten parannusta 3-6-grammien käyttäminen antoi romanian kielellä (0,59 ilman alimerkkijonoja, 0,66 3-6-grammien kanssa) ja arabialla (0,51 ilman alimerkkijonoja, 0,55 3-6-grammien kanssa), mutta myös espanjalla korrelaatio kasvoi (0,57 ilman alimerkkijonoja, 0,59 3-6-grammien kanssa). Suomen kielellä korrelaation parantuminen on samaa luokkaa arabian kielen kanssa: 3-6-grammeja hyödyntävän ja niitä hyödyntämättömän mallin erotus oli suomella 0,034, arabialla 0,04. Valitsemalla sopivat n-grammit voidaan saada kuitenkin vielä selvästi suurempi parannus (parhaan ja alimerkkijonoja hyödyntämättömän mallin korrelaatioiden erotus 0,072).

Erikielisten mallien vertailu on kuitenkin parhaimmillaamkin suuntaa-antavaa, sillä datasettien kääntämiseen liittyvien kysymysten lisäksi eri kielet on opetettu eri aineistoilla. Kaikissa Bojanowskin ym. kokeissa kuten tämänkin tutkielman kokeissa opetusaineistona on toiminut kyseisen kielen Wikipedia-aineisto. Esimerkiksi englanninkielinen Wikipedia on kuitenkin huomattavasti laajempi kuin suomenkielinen Wikipedia, joten sillä opetetut mallit voisi odottaakin toimivan suomenkielisiä malleja paremmin. Kuitenkaan opetusaineiston kokoero ei tee vertailuista mahdottomia. Bojanowski ym. testasivat asiaa englanninkielisellä harvinaisten sanojen RW-datasetillä [24], jossa sanat on valittu esiintymistiheyden perusteella ja joka sisältää myös taipuneessa muodossa olevia sanoja. Sillä saadaan CBOW-mallilla jopa parempi korrelaatiotulos (0,45) käyttämällä vain 1% Wikipedia-aineistosta kuin käyttämällä koko aineistoa (korrelaatio 0,43). Mallin laatu ei siis välttämättä kasva opetusaineiston määrän kasvaessa.

Aiempaan suomen kieltä koskevaan kirjallisuuteen verrattuna korrelaatiot ovat hyviä, mutta eivät poikkeuksellisen hyviä. Venekoski ym. [42] tekivät muiden kokeittensa joukossa myös kokeen, jolla testattiin skip-gramia hyödyntävän fastTextin toimivuutta samankaltaisuustehtävässä, ja opetusaineistona oli suomenkielinen Wikipedia, kuten tässä tutkielmassa. He saivat tulokseksi 0,212, kun $d = 100$ ja käytössä olivat 3-6-grammit. Tässä tutkielmassa käytettiin samaa menetelmää ja vastaavalla aineistoa.

Samoilla parametreilla saatiin tässä tutkielmassa melkein sama tulos, 0,210. Eri n-grammien pituuksia kokeilemalla saatiin parhaiksi tuloksiksi Venekosken ym. kääntämällä 0,271 SimLex-datasetillä (6-6-grammit). Heidän tuottamallaan FinnSim-datasetillä tulos oli 0,259 (4-9-grammit). Vastaavasti samoilla n-grammeja koskevilla parametreilla, mutta ulottuvuuksien määrää

vaihtamalla, saatiin parhaiksi tuloksiksi FinnSim-testeissä 0,254 ($d = 600$) ja SimLex-testeissä 0,274 ($d = 600$).

Kokeiden tulos näillä dataseteillä saatiin käyttämällä 4-7-grammeja ja moniulotteisia vektoreita: FinnSim-datasetin korrelaatio on parhaimmillaan 0,287 ($d = 500$) ja SimLex-datasetin 0,297 ($d = 600$).

Venekoski ym. [42] saavat Wikipedia-datasetillä parhaimmillaan samankaltaisuustulokseksi 0,278 käyttämällä CBOW-menetelmää fastText-mallissa skip-gramin sijaan. Tulos on siis hiukan huonompi kuin paras tässä tutkielmassa löydetty yhdistelmä.

Tätä tutkielmaa varten yritettiin toistaa Venekosken ym. tekemää koetta, jossa käytettiin CBOW-fastText-mallia. Kuitenkaan tutkielmaa tehdessä ei onnistuttu toistamaan tulosta, vaan CBOW-mallin korrelaatioksi SimLex-datasetillä saatiin 0,196, kun $d = 100$ ja käytössä ovat 3-6-grammit. Skip-gram-mallilla vastaava luku on 0,210. CBOW-mallin korrelaatioksi FinnSim-datasetillä saatiin puolestaan 0,188, kun $d = 100$ ja käytössä ovat 3-6-grammit. Skip-gram-mallilla vastaava luku on 0,191.

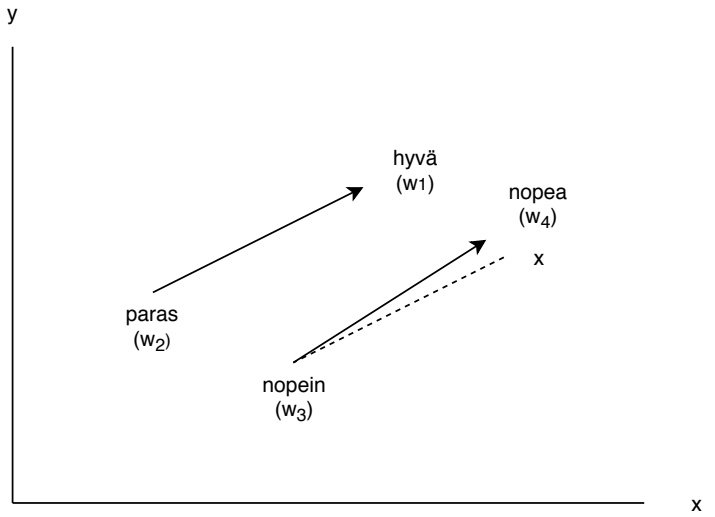
Venekoski ym. tutkimuksesta ei selviä, mitkä parametrit olivat tarkalleen käytössä, joten on mahdollista, että kokeen toistaminen ei ole onnistunut oikein. Kuitenkin $d = 100$ ja 3-6-grammit ovat fastText-kirjaston oletusparametrit, joten ne ovat todennäköinen vaihtoehto. Tutkielmaa varten kokeiltiin myös muita parametreja, mutta millään tarkastellulla yhdistelmällä CBOW-malli ei pärjännyt skip-gram-mallia paremmin.

On vaikea sanoa, miksi CBOW-mallin tulos on tässä tutkielmassa huonompi kuin skip-gram-mallin, vaikka Venekoski ym. saivat päinvastaisen tuloksen. Yksi selitys voi olla käytetyssä datasetissä: vaikka nämä luvut olivat tulosta Wikipedia-datasetistä, ovat Venekoski ym. käyttäneet eri päivänä ladattua ja eri tavalla esiprosessoitua aineistoa kuin tässä tutkielmassa. Ainakin osa eroista johtuu siis todennäköisesti opetusdatasta.

4 Analogiatestit

Mallien laatua voidaan mitata myös tutkimalla, miten hyvin ne suoriutuvat analogiaparien tunnistamisessa. Mikolov ym. [26] esittävät, että word2vec-malleilla muodostetussa jatkuvassa sana-avaruudessa sanojen välisiä suhteita ilmentää muukin kuin niiden sijaitseminen lähellä toisiaan. Koska sanat esitetään vektoreina, voidaan niillä tehdä myös yhteen- ja vähennyslaskutoimituksia kuten muillakin vektoreilla. Word embedding -vektoreilla tehdyillä laskutoimituksilla voidaan tehdä semanttisia vertailuja, koska laskutoimitukset vastaavat sanojen välisiä semanttisia suhteita.

Ihmiselle on esimerkiksi selvää, että sana *pieni* on samanlaisessa suhteessa sanaan *pienempi* kuin sana *suuri* on sanaan *suurempi*. Siksi ihminen kykenee myös vastaamaan kysymykseen siitä, mikä sana on vastaa sanalle *pieni* sitä, mitä *suurempi* vastaa sanalle *suuri*.



Kuva 4: Sanojen väliset analogiasuhteet vektoriavaruudessa. Sanan *paras* suhde sanaan *hyvä* on lähellä sanan *nopein* suhdetta sanaan *nopea*. Sanaa *nopea* vastaava vektori w_4 on lähellä vektoria $x = w_1 - w_2 + w_3$.

Tällaista suhdetta kutsutaan analogiaksi. Sanojen *suuri* ja *suurempi* välinen analogia on luonteeltaan taivutusmuotoihin perustuva, sillä *suurempi* on adjektiivin *suuri* komparatiivimuoto. Word embedding -vektorit ilmentävät kuitenkin myös puhtaasti semanttisia analogiasuhteita. Esimerkiksi *äiti* ja *isä* ovat keskenään samanlaisessa semanttisessa suhteessa kuin *tytär* ja *poika*, sillä *isä* on miespuolinen vastine sanalle *äiti* ja *poika* miespuolinen vastine sanalle *tytär*.

4.1 Datasetit

Sana-analogiakyselyitä sisältävillä dataseteillä voidaan tutkia analogioiden löytymistä sana-avaruudesta. Bojanowskin ym. [4] käyttämässä Mikolovin ym. [26] kehittämässä menetelmässä valitaan ensin kaksi sanaparia, joiden sanojen välillä on samanlainen analogiasuhde. Merkitään sanaparien sanoja vastaavia vektoreita $\{w_1, w_2\}$ sekä $\{w_3, w_4\}$. Tehdään laskutoimitus $w_1 - w_2 + w_3 = x$. Tutkitaan avaruutta vektorin x ympäriltä ja etsitään sitä lähin avaruudesta löytyvä sanavektori w_4'. Mikäli $w_4' = w_4$, malli toteuttaa halutun analogian. Joskus voi käydä niin, että w_4' on sama sana kuin jokin kyselyn muista sanoista w_1, w_2 tai w_3. Tällöin w_4' hylätään: sen sijaan valitaan sellainen w_4'', että se on seuraavaksi lähimpänä vektoria x ja tutkitaan, onko $w_4'' = w_4$.

Tässä tutkielmassa analogiakyselyt on toteutettu Mikolovin ym. [26] käyttämällä top1-menetelmällä, jossa vaaditaan, että w_4' on kaikista lähin sanavektori x:n nähden, jollei se kuulu muihin kyselyn sanoihin. Tutkielmassa

Taulukko 3: Esimerkkejä sanojen merkitykseen ja muotoon liittyvistä suhteista käännetyssä SSWR-fi-datasetissä. Ylempänä olevat analogiasuhteet liittyvät sanan merkitykseen, alemmat ovat johdannaisia tai taivutusmuotoja. Suomentamisen yhteydessä datasettiä lyhennettiin jättämällä pois sellaiset muotokyselyt, jotka eivät kääntyneet luontevasti suomeksi.

Suhde	Esimerkkipari	
Maa ja pääkaupunki (Yleiset)	Kairo	Egypti
Maa ja valuutta	Ruotsi	kruunu
Kaupunki ja osavaltio	Orlando	Florida
Mies ja nainen	sulhanen	morsian
Adjektiivi ja adverbi	ilmeinen	ilmeisesti
Vastakohta	rehellinen	epärehellinen
Kompataritiivi	kova	kovempi
Superlatiivi	pitkä	pisin
Kansalaisuus	Intia	intialainen
Substantiivin monikko	lehmä	lehmät
Verbin 3. persoona	mennä	menee

on tehty kokeita top1-menetelmän lisäksi top3-menetelmällä, jossa etsitään kolme x:ää lähinnä olevaa vektoria w_4', w_4'' ja w_4''', missä vektorit eivät myöskään saa olla samoja kuin kyselyn muut vektorit. Jos $w_4 \in \{w_4', w_4'', w_4'''\}$, niin malli toteuttaa halutun top3-analogian. Siis top3-kyselyt ovat top1-kyselyitä helpompia.

4.1.1 SSWR ja SSWR-fi

Yksi myös Bojanowskin ym. [4] hyödyntämistä arviointiresursseista on Mikolovin ym. [26] luoma analogiakyselyitä sisältävä datasetti Semantic-Syntactic Word Relationship test set eli SSWR. Datasetissä esitetyt semanttisten analogioiden sanaparit liittyivät maantietoon (**maa ja pääkaupunki**, **maa ja valuutta**, **osavaltio ja kaupunki**) ja ihmisiin (**mies ja nainen**). Parien sanat ovat pääosin perusmuotoisia substantiiveja. Sananmuotoihin ja johdannaisiin liittyvät kysymykset taas kattavat laajemmin eri sanaluokkia. Esimerkkejä datasetin sanoista on taulukossa 3.

SSWR-datasetti on jaettu kahteen osaan: ensimmäisessä puolikkaassa ("semantic") on sanojen merkitykseen liittyviä kyselyitä, ja toisen puoliskon ("syntactic") kyselyt liittyvät sanojen muotoon, eli analogiaparin toisena tulevat sanat on saatu johtamalla tai taivuttamalla ensimmäiseksi tulevista sanoista. Tutkielmassa käytetystä termistöstä on kerrottu lisää luvussa 1.1.

Tutkielmaa varten SSWR-datasetti käännettiin suomeksi. Tähän käännettyyn ja kääntyvyyssyistä lyhennettyyn datasettiin viitataan jatkossa nimellä

SSWR-fi. [3]

SSWR-datasettiä on käännetty aiemminkin muille kielille englannista. Köper ym. [22] käänsivät datasetin saksaksi. Käännöksessä jätettiin pois suhde **adjektiivi ja adverbi**, sillä suhdetta ei ole olemassa saksan kielessä. Vastaavasti Berandi ym. [3] tuottivat SSWR-datasetille italiankielisen käännöksen. Myös italiankielisessä käännöksessä jouduttiin jättämään joitakin sanoja pois muun muassa siksi, että komparatiivin muodostuksessa käytettiin joidenkin sanaston sanojen kohdalla sanaliittoa yhden sanan taivuttamisen sijaan. Sanavektorit muodostetaan yksittäisten sanojen, ei sanaliittojen tai fraasien perusteella, joten word embedding -esitysten tutkimiseen tämä taivutusmuodot eivät olisi soveltuneet. Berandi ym. tekivät datasettiin myös lisäyksiä italian verbien ja nominien taivutukseen sekä italialaiseen paikannimistöön liittyen.

Joillekin kielille on SSWR-datasetin sijaan tehty sen pohjalta omankielinen uusi datasetti. Näin tekivät esimerkiksi Svoboda ym. [40] luoden tšekinkielisen datasetin. Osa datasetin kyselyistä on samoja kuin SSWR:ssä, osa taas tšekinkielen ominaisuuksien perusteella kehitettyjä. Esimerkiksi eroa maskuliinisten ja feminiinisten kansalaisuutta ja ammattia koskevien sanojen välillä ei ole englannissa, mutta slaavilaisista kielistä se löytyy.

Suomenkielisessä, tätä tutkielmaa varten tehdyssä käännöksessä jätettiin pois kategoriat **present participle** ja **past tense**, sillä niissä sanaparin toinen osa on ollut englannin verbien partisiippimuodossa, jonka merkitys on hyvin erilainen kuin mikään suomen verbien partisiippimuoto. Esimerkiksi *code* ja *coding* on luonteeltaan ja esiintymistiheydeltään erilainen sanapari kuin *koodata* ja *koodaamassa* tai *koodaava*, joten vertailu ei ole mielekäs.

Käännöksessä on jätetty pois myös joitakin sanapareja yksittäisistä kategorioista. Joillekin pareille ei löytynyt mielekästä käännösvastiketta (esimerkiksi *policeman* ja *policewoman*, *him* ja *her*, *most* ja *mostly* sekä *possibly* ja *impossibly*). Joillekin sanoille ei löytynyt käännösvastinetta, koska prototyyppinen käännösvastine esiintyi datasetin käännöksessä jo toisen sanan käännöksenä. Esimerkiksi adjektiivien joukossa esiintyivät sekä sanat *big* että *large* sekä *long*, *high* ja *tall* ja verbien joukossa *speak* ja *talk*. Verbin 3. persoona -kategoriasta on poistettu sanat, joiden suomenkielinen käännösvastine on sama lekseemille ja sananmuodolle, kuten *laulaa* ja *löytää*.

Käännöksessä on pyritty valitsemaan mahdollisimman prototyyppinen käännösvastine, koska alkuperäisenkin datasetin sanat ovat prototyyppisiä ja yleisiä sanoja. Joissakin käännöksissä vastineen merkitys on rajallisempi kuin lähtökielen sanan, sillä esimerkiksi sanaa *niece* ei voida suomessa ilmaista ottamatta kantaa siihen, onko kyseessä siskontyttö vai veljentyttö. Tällaiset sanat pidettiin datasetissä ja niille valittiin merkitykseltään kapeampi käännösvastine, koska suhde sanojen *siskontyttö* ja *siskonpoika* on samanlainen huolimatta siitä, muodostuuko sukulaisuussuhde veljen vai siskon kautta.

[3] Käännetty datasetti löytyy osoitteesta https://github.com/adaisti/fin-eval.

Datasetti on pyritty pitämään mahdollisimman lähellä alkuperäistä, vaikka sen sisältämä tosimaailman tieto olisi ehtinyt muuttua. Esimerkiksi Liettuan ja Latvian valuutat on jätetty latiksi ja litiksi, vaikka tutkielman suomenkielisten mallien opetuksessa käytetty Wikipedia-datasetti on vuodelta 2017, jolloin maissa on ollut jo käytössä euro. Tämä tehtiin siksi, että myös Bojanowskin ym. [4] kokeet tehtiin vastaavasti vanhentunutta tietoa sisältävällä datasetillä.

Käännösratkaisujen takia tutkielmassa käytetty SSWR-fi-datasetti on lyhyempi kuin alkuperäinen SSWR. Merkitykseen liittyviä kategorioita on lyhennetyssä suomennetussa datasetissä 5, muotoon liittyviä 7. Suomennettu SSWR-fi-datasetti sisältää 15344 kyselyä, joista 8743 liittyy merkitykseen, 7301, muotoon. Erillisiä sanapareja näissä kyselyissä on yhteensä 471, joista 257 liittyy merkitykseen, 214 muotoon.

SSWR-datasetissä kaikista mahdollisista sanaparien välisistä permutaatioista ei tehdä kyselyitä, vaan Mikolov ym. valitsivat satunnaisesti kyselyissä rinnastetut parit [26]. Tässä tutkielmassa on käytetty samoja pareja kuin alkuperäisessä datasetissä. Datasetti on siis mahdollisesti painottunut joidenkin kyselyiden osalta. Vaikka kaikkia permutaatioita voisi pitää kiinnostavina, on tässä tutkielmassa kuitenkin otettu huomioon vain alkuperäisessä SSWR-datasetissä esitetyt vertailut, jotta luvut olisivat mahdollisimman vertailukelpoisia muuhun tutkimukseen nähden.

Datasettien kokoa tarkkaillessa on kuitenkin hyvä ottaa huomioon, että vain osa niiden sisältämistä kyselyistä löytyi fastText-mallin tuottamasta sana-avaruudesta. Bojanowskin ym. [4] esimerkin mukaisesti datasettejä koskevissa tuloksissa huomioidaankin ainoastaan ne kyselyt, joiden kohdalla sanaparien kaikki neljä sanaa löytyivät sana-avaruudesta. Kyselyissä on huomioitu 735 000 yleisintä sanaa, sillä sitä useampien sanojen mukaan ottaminen ei enää kasvattanut hyödynnettyjen kyselyiden määrää.

Sellaisia kyselyitä, joissa jokin sana ei löydy avaruudesta, on verrattain paljon, vaikka mahdollisimman moni kysely yritettiin saada mukaan. Nyt hyödynnettyjä kyselyitä on 11012, mikä on 71,77% kaikista SSWR-fi-datasetin kyselyistä.

4.1.2 FinnAn

Tutkielmassa on hyödynnetty myös Venekoski ym. [42] julkaisemaa analogiadatasettiä, jossa analogioita tutkitaan seitsemän suhteen avulla: **adjektiivi ja vastakohta** (esim. *leveä, kapea*), **pääkaupunki ja maa** (esim. *Tallinna, Viro*), **luku ja järjestysluku** (esim. *yksi, ensimmäinen*), **maa ja valuutta** (esim. *Japani, jeni*), **nainen ja mies** (esim. *naisellinen, miehekäs*), **jääkiekkojoukkue ja kaupunki** (esim. *Tappara, Tampere*) sekä **suunta ja vastakkainen suunta** (esim. *ylös, alas*). Jäljempänä tästä datasetistä käytetään nimeä FinnAn-datasetti.

FinnAn-datasetti on luotu suoraan suomen kielelle, joten se huomioi

SSWR-fi-datasettiä paremmin kulttuurikohtaiset erityispiirteet. Esimerkiksi suhde **jääkiekkojoukkue, kaupunki** on Suomessa kulttuurisesti merkittävä ja koskee nimenomaan suomalaisia kaupunkeja. Tätä suhdetta tuskin löytyisi muunkielisen aineiston pohjalta muodostetusta sana-avaruudesta kovin selkeästi.

FinnAn-datasetti sisältää osittain samoja suhteita kuin SSWR-fi-datasetti, kuten **nainen ja mies**, sillä se on kehitetty SSWR-datasetin pohjalta. Toisin kuin SSWR-fi-datasetissä, on FinnAn-datasetissä myös samassa taipuneessa muodossa olevien sanojen muodostamia sanapareja, kuten pari *ylhäällä, alhaalla*. Datasetissä tutkitaan laajemmin myös muiden kuin substantiivien välisiä semanttisia suhteita.

FinnAn-datasetti ei sisällä sanojen taipumiseen liittyviä analogioita, vaan lähes pelkästään sanan merkityksen analogioita. Sen ainoa sananmuodostukseen liittyvä analogiaosio tutkii suhdetta **luku ja järjestysluku**. Eli vaikka datasetti sisältääkin sanoja taipuneissa muodoissa, se on luonteeltaan ensisijaisesti semanttisia analogioita tutkiva, eikä tutki taivutuksen analogioita lainkaan.

FinnAn-datasetti on selvästi pienempi kuin SSWR-fi-datasetti: siinä on vain 1037 kyselyä, jotka muodostuvat 85 eri sanaparista. Kyselyistä ainoastaan yksi kategoria koskee sananjohdannaisia, ja siinä on 109 kyselyä ja 9 eri sanaparia. Kyselyistä 603 oli sellaisia, että kaikki neljä sanaparien sanaa löytyivät sana-avaruudesta. Tämä on 58,63% kyselyistä, eli vähemmän kuin SSWR-fi-datasetin kohdalla. Kuten aiemmin, otettiin oikeudellisuutta tarkasteltaessa huomioon vain ne kyselyt, joiden kaikki sanat löytyivät sana-avaruudesta.

Mikolovin ja Venekosken dataseteille on yhteistä se, että ne sisältävät analogiakyselyitä kumpaankin suuntaan, siis esimerkiksi sekä kyselyn *äiti - tyttö + poika = isä* että *tyttö - äiti + isä = poika*. Kyselyt kumpaankin suuntaan ovat tarpeellisia, sillä voi olla, että johonkin suuntaan toteutettuna kysely saa ensimmäiseksi hakutulokseksi halutun sanan, mutta toiseen suuntaan saatu sana ei ole kyselyn hakema. Ideaalitapauksessa suhteet olisivat niin lähellä toisiaan, että ei ole väliä, kummin päin kyselyn toteuttaa. Toisaalta kyselyissä on aina yksi oikea vastaus, vaikka tosimaailmassa oikeita vastauksia voisi olla enemmänkin. Esimerkiksi sanan *poika* naispuolinen vastine voisi olla joko *tyttö* tai *tytär* riippuen siitä, tarkoitetaanko sillä perhesuhdetta vai nuorta henkilöä. Kuitenkin kyselyn *tyttö - äiti + isä* tulos tulisi yksikäsitteisesti olla *poika*. Siis joillakin kyselyillä suunnilla on väliä sanojen merkitysten vuoksi, koska kyselyn tulos saattaa olla oikea vain toiseen suuntaan.

Yksikäsitteisyyden puute on yksi syy siihen, miksi tässä tutkielmassa on tehty kokeita paitsi top1-meneltemällä, myös top3-menetelmällä, jossa riittää, että oikea vastaus löytyy mallin antamien vastausten kolmen parhaan joukosta.

4.2 Tulokset

Tutkielmaa varten testattiin mallien oikeellisuutta edellisessä luvussa kuvailtujen analogiakyselyiden avulla. Kyselyiden oikeellisuutta tutkittiin semanttisen samankaltaisuuden tapaan sekä erikokoisilla sanavektorin ulottuvuuksilla että eri n-grammien pituusväleillä. Tarkoituksena oli selvittää, parantaako n-grammien käyttö mallien kykyä selviytyä kaikenlaisista analogiakyselyistä ja parantuuko tai huonontuuko mallin laatu eri verran merkitykseen ja muotoon liittyvissä kyselyissä.

Kuten semanttisen samankaltaisuuden kohdalla, kokeissa muodostettiin kielimalleja fastText-kirjaston avulla eri parametreilla. Parametrit pyrittiin käymään läpi mahdollisimman laajasti, joten kokeita on tehty useammalla eri yhdistelmällä kuin Bojanowski ym. [4] tekemissä kokeissa.

Tulosten perustella käy ilmi, että alimerkkijonojen hyödyntäinen lisää mallin kykyä vastata muotoa koskeviin analogiakyselyihin, mutta samaan aikaan se laskee mallin tarkkuutta merkitykseen liittyvissä analogiakyselyissä. Ilmiö on nähtävissä sekä eriulotteisilla vektoreilla että erilaisilla n-grammien pituuksilla eikä siihen vaikuta se, vaaditaanko kyselyiltä oikeaa vastausta ensimmäisenä, vai riittääkö, että oikea vastaus on kolmen parhaan vastauksen joukossa.

4.2.1 Vektorin ulottuvuuksien määrä

Word embedding -mallia opetettaessa pitää valita, kuinka suuriulotteisia vektoriesityksiä sanoille halutaan muodostaa. Kuten aiemmin, vektoriesityksen ulottuvuuksien määrä on d eli verkon piilokerroksen koko.

Tässä tutkielmassa malleja opetettiin kokeilemalla d:lle eri arvoja joukosta $\{100, 200, ..., 700\}$. Ulottuvuuden vaikutusta testattiin menetelmällä, jossa alimerkkijonoja ei hyödynnetty (merkitään 0-grammilla) sekä 3-6-grammeilla ja 4-7-grammeilla. Tulokset on esitetty taulukoissa 4 - 5.

Kuten taulukosta 4 havaitaan, SSWR-fi-datasetillä top1-kyselyt toteutuvat useammin oikein, kun d on varsin pieni. Kaikentyyppisiä kyselyitä tarkastellessa parhaat tulokset on saatu valitsemalla $d = 100$, jos n-grammeja hyödynnetään, tai $d = 200$, jos ne eivät ole käytössä. Paras tulos, 36,24% on saatu valitsemalla $d = 100$ ja 4-7-grammit.

Kun tarkastellaan pelkästään merkityskyselyitä, tulokset käyttäytyvät d:n suhteen samalla tavalla. Pienet d:n arvot antavat parhaat tulokset, ja paras valinta on ottaa $d = 200$ eikä alimerkkijonoja, jolloin muotokyselyistä saadaan oikein 31,11%.

Merkityskyselyiden kohdalla on huomionarvoista myös, että kun d on melko pieni, eroa n-grammeja hyödyntävillä ja niitä käyttämättömillä malleilla ei ole merkittävästi. Kuitenkin d:n kasvattaminen laskee n-grammeja hyödyntävien mallien laatua voimakkaammin kuin alimerkkijonoja käyttämättömien mallien. Esimerkiksi kun $d = 100$, ilman alimerkkijonoja malli saa

Taulukko 4: SSWR-fi-datasetin analogiakyselyiden tulokset erilaisilla sanavektorin ulottuvuuksilla ja n-grammeilla. Vasemmassa sarakeryhmässä on kaikkien kyselyiden oikeellisuusprosentti. Keskimmäisessä sarakeryhmässä on merkityskyselyiden oikeellisuusprosentti, oikeassa muotokyselyiden. Lihavoitu arvo on paras sarakeryhmittäin ja riveittäin.

	Yhteensä			Merkitykset			Muodot		
	n-grammi			n-grammi			n-grammi		
d	0	3 - 6	4 - 7	0	3 - 6	4 - 7	0	3 - 6	4 - 7
100	30,60	35,66	**36,24**	**27,67**	25,36	26,12	36,48	56,32	**56,53**
200	33,71	34,53	**34,73**	**31,11**	21,11	21,67	38,91	**61,42**	60,90
300	**31,37**	30,49	29,83	**28,92**	14,20	14,84	36,26	**63,17**	59,89
400	**30,22**	26,73	25,49	**27,73**	8,77	9,49	35,23	**62,76**	57,57
500	**28,80**	23,95	22,49	**27,00**	5,49	6,12	32,39	**60,95**	55,31
600	**26,67**	21,78	20,36	**24,85**	3,44	4,33	30,31	**58,53**	52,50
700	**25,44**	20,82	18,37	**23,93**	2,52	2,64	28,46	**57,52**	49,90

Taulukko 5: FinnAn-datasetin analogiakyselyiden tulokset erilaisilla sanavektorin ulottuvuuksilla ja n-grammeilla. Vasemmassa sarakkeessa on kaikkien kyselyiden oikeellisuusprosentti. Keskimmäisessä sarakkeessa merkityskyselyiden oikeellisuusprosentti, oikeassa muotokyselyiden. Taulukkoa tarkasteltaessa on huomioitava, että muotokyselyitä on datasetissä vain vähän, joten niitä koskevat arvot ovat mukana lähinnä kattavuuden, ei merkittävyyden vuoksi.

	Yhteensä			Merkitykset			Muodot		
	n-grammi			n-grammi			n-grammi		
d	0	3 - 6	4 - 7	0	3 - 6	4 - 7	0	3 - 6	4 - 7
100	**52.80**	52.14	52.14	**55.83**	55.65	55.30	**11.90**	4.76	9.52
200	**54.11**	49.67	49.67	**57.07**	52.12	52.12	14.29	**16.67**	**16.67**
300	**55.92**	38.49	40.95	**59.36**	40.46	43.11	9.52	**11.90**	**11.90**
400	**55.76**	29.11	28.95	**59.36**	30.21	30.04	7.14	**14.29**	**14.29**
500	**53.45**	21.38	22.86	**56.89**	22.08	23.32	7.14	11.90	**16.67**
600	**50.16**	14.47	15.30	**53.18**	14.66	15.90	9.52	**11.90**	7.14
700	**49.01**	11.51	11.18	**51.94**	11.48	11.13	9.52	**11.90**	**11.90**

Taulukko 6: SSWR-fi-datasetin top3-analogiakyselyiden oikeellisuusprosentit. Tässä taulukossa analogiakyselyt on toteutettu niin, että riittää, että kyselyn oikea vastaus löytyy kolmen parhaan ehdotuksen joukosta.

	Yhteensä			Merkitykset			Muodot		
	n-grammi			n-grammi			n-grammi		
d	0	3 - 6	4 - 7	0	3 - 6	4 - 7	0	3 - 6	4 - 7
100	44.01	49.39	**49.94**	42.52	40.14	**40.53**	46.98	67.94	**68.79**
200	46.49	**50.65**	50.32	**45.01**	38.19	37.85	49.47	**75.63**	75.31
300	44.26	**45.73**	**45.69**	42.83	29.97	31.31	47.12	**77.33**	74.52
400	**42.46**	39.74	39.58	**40.52**	20.66	23.19	46.36	**77.98**	72.41
500	**39.97**	35.18	34.04	**38.90**	14.94	16.06	42.10	**75.74**	70.10
600	**37.79**	31.37	30.19	**36.52**	10.40	11.88	40.33	**73.42**	66.88
700	**36.71**	28.29	27.02	**35.78**	6.87	8.14	38.55	**71.21**	64.86

oikein 27,67% kyselyistä, mikä on vain hiukan enemmän kuin 3-6-grammien 25,35%. Kuitenkin kun $d = 700$, ilman alimerkkijonoja tulos on 23,93%, mutta 3-6-grammeilla vain 2,52%.

Muotoa koskevia kyselyitä tarkastellessa hiukan isommat d:n arvot antavat parempia tuloksia, ja paras tulos on saatu valitsemalla $d = 300$ sekä 3-6-grammit. Kuitenkin myös muotokyselyiden kohdalla suuret d:n arvot haittaavat mallin toimintaa. Muotoa koskevissa kysymyksissä alimerkkijonoja hyödyntävät mallit pärjäävät selvästi niitä käyttämättömiä malleja paremmin. Kuitenkaan alimerkkijonoja käyttämättömät mallit eivät huonone yhtä merkittävästi d:n kasvaessa kuin niitä hyödyntävät mallit merkityskyselyiden kohdalla. Esimerkiksi kun $d = 100$, saadaan ilman alimerkkijonoja kyselyistä oikein 36,48%, ja kun $d = 700$, oikeita tuloksia on 28,46%.

FinnAn-datasetillä (taulukossa 5) tulokset noudattelevat keskinäisiltä suhteiltaan SSWR-fi-datasettiä koskevia tuloksia. Datasetissä on vain yksi muotoa koskeva kyselytyyppi, joten se vertautuu kokonaisuutena enemmänkin SSWR-fi-datasetin merkityskyselyihin kuin koko SSWR-fi-datasettiin, jossa muotokyselyitä on merkityskyselyitä enemmän. Huomataan kuitenkin, että FinnAn-datasetillä merkityskyselyistä on saatu selvästi parempia tuloksia kuin SSWR-fi-datasetillä, ja sen muotokyselyiden tulokset ovat taas huonompia kuin SSWR-fi-datasetin.

FinnAn-datasetillä paras tulos on saatu valitsemalla $d = 300$, eli toisin kuin SSWR-fi:llä, pienimmät mahdolliset d:n arvot eivät ole parhaita. Kuitenkin erot eri d:n arvoilla ovat pieniä. Jälleen havaitaan, että ensin tulos paranee ja lähtee sitten huononemaan, kun d:tä kasvatetaan. Siis d:lle löytyy jälleen sellainen arvo tai väli, jolla tulos on optimaalinen.

Myös SSWR-fi-datasetillä tehdyt top3-analogiakyselyt ovat linjassa top1-analogiakyselyiden antamien tulosten kanssa. Top3-kyselyiden tulokset on esitetty taulukossa 6. Kuten top1-kyselyissä, kummankin tyylisiä kyselyitä tarkasteltaessa parhaat d:n arvot ovat pieniä, 100 tai 200.

Tässäkin kyselytyypissä paras d:n arvo vaikuttaisi olevan merkityskyselyillä 200 ja muotokyselyillä hiukan suurempi, 300 tai 400. Vaikka top3-kyselyt ovat helpompia kuin top1-kyselyt, huonontaa d:n kasvattaminen edelleen merkityskyselyissä 3-6-grammien ja 4-7-grammien tuottamia tuloksia paljon enemmän kuin alimerkkijonoja hyödyntämättömien mallien. Samoin muotokyselyissä n-grammeja käyttämättömien mallien tulokset huonontuivat isoilla d:n arvoilla enemmän kuin 3-6-grammien ja 4-7-grammien mallit.

FinnAn-datasetille ei toteutettu tässä tutkielmassa top3-kyselyitä. Datasetti oli pienempi kuin SSWR-fi ja sisälsi lähinnä merkitystä koskevia kyselyitä. Tutkielmassa valittiin keskittyä SSWR-fi-datasetin kahden eri kyselytyypin tutkimiseen, koska niitä vertailtaessa ilmeni mielenkiintoisia ilmiöitä.

4.2.2 N-grammien pituus

Kuten edellisessä kappaleessa tutkituista taulukoista 4 - 6 havaitaan, kyselyjen tyypillä on paljon merkitystä sen kannalta, hyödyttääkö vai haittaako alimerkkijonojen käyttö niiden onnistumista. Jos kysely koskee sanan merkitystä, alimerkkijonoista vaikuttaa olevan vain haittaa. Viime luvussa nähtiin, että haitta korostuu erityisesti, jos d on suuri. Jos taas kysely koskee sanan muotoa, alimerkkijonoja hyödyntävät mallit suoriutuvat niistä selvästi n-grammeja käyttämättömiä malleja paremmin riippumatta siitä, onko n-grammeiksi valittu 3-6-grammit vai 4-7-grammit.

Taulukoista huomataan myös, että yleisesti ottaen 3-6-grammit pärjäsivät hiukan 4-7-grammeja paremmin muotokyselyissä. Tässä luvussa tutkitaan, millaiset n-grammien välit toimivat parhaiten merkityskyselyissä, muotokyselyissä sekä kummankin tyylisissä kyselyissä. Välejä on testattu kiinteällä d:n arvolla, eli on valittu $d = 300$. Valinta on sama kuin Bojanowskin ym. [4] tekemissä kokeissa, ja myös Mikolov ym. [26] käyttivät sitä yhtenä datasettiä esittelevänä perusulottuvuutena. Aiemmassa luvussa havaittiin, että kun $d = 300$, muotokyselyissä on usein saatu n-grammeja hyödyntävillä malleilla paras tai toiseksi paras tulos. Kutenkin merkityskyselyissä n-grammien antamat tulokset olivat selvästi huonompia, kun $d = 300$ (esimerkiksi SSWR-fi top1-kyselyiden 3-6-grammien 14,20%) kuin jos $d = 200$ (SSWR-fi top1-kyselyiden 3-6-grammeilla 21,11%). Siis valitsemalla $d = 300$ saattavat erot huonojen ja hyvin n-grammivalintojen välillä korostua entisestään. Tämän luvun kannalta se on hyvä, koska tarkoituksena on selvittää, millaiset n-grammit toimivat parhaiten.

Tarkastellaan aluksi SSWR-fi-datasetin top1-kyselyitä. Tulokset on esitelty taulukoissa 7a - 7c. Taulukoissa on korostettu jokaisen sarakkeen paras arvo. Siis lihavoitu arvo kertoo, mikä on paras valinta i:ksi, kun j on kiinnitetty ja kun käytetyt n-grammit ovat välillä $[i, j]$. Esimerkkinä taulukosta 7a nähdään, että jos valittu n-grammi saa olla suurimmillaan 6-grammi, niin väli kannattaa valita alkavaksi 2-grammeista, koska silloin mallin tarkkuus

Taulukko 7: SSWR-fi-datasetin analogiakyselyiden oikeellisuusprosentti eri n-grammien pituuksilla: i on minimipituus, j maksimi.

(a) Kaikki kyselyt.

$i\backslash j$	2	3	4	5	6	7	8	9	10
2	**29,03**	28,59	29,48	29,80	**30,95**	**31,57**	31,98	**31,95**	**31,66**
3		**30,00**	**29,87**	**29,95**	30,49	31,27	**32,04**	31,52	31,39
4			29.79	29.29	29.38	29.83	30.38	30.50	30.78
5				29.94	28.79	28.44	30.32	30.30	29.02
6					29.05	29.02	29.68	29.09	29.49
7						27.98	28.02	29.49	29.43
8							28.94	28.68	29.22
9								28.92	29.48
10									30.32

(b) Merkitystä koskevat kyselyt.

$i\backslash j$	2	3	4	5	6	7	8	9	10
2	**16.28**	11.19	12.21	13.47	15.20	17.27	17.64	17.80	17.59
3		**13.77**	12.26	12.73	14.20	15.63	16.92	16.50	16.41
4			**13.95**	12.93	13.92	14.84	15.86	16.35	17.67
5				**15.91**	15.05	15.09	17.54	17.87	17.46
6					**18.01**	18.24	19.61	20.28	20.58
7						**19.41**	20.25	22.02	22.58
8							**22.89**	23.34	24.72
9								**25.29**	25.66
10									**27.74**

(c) Muotoa koskevat kyselyt.

$i\backslash j$	2	3	4	5	6	7	8	9	10
2	**54.60**	**63.47**	64.09	62.54	62.51	60.22	60.74	60.30	59.86
3		62.54	**65.16**	**64.47**	**63.17**	**62.65**	**62.35**	**61.64**	**61.42**
4			61.56	62.07	60.35	59.89	59.48	58.88	57.08
5				58.06	56.32	55.20	55.93	55.23	52.20
6					51.19	50.64	49.85	46.74	47.34
7						45.16	43.60	44.45	43.17
8							41.06	39.37	38.25
9								36.21	37.14
10									35.50

Taulukko 8: FinnAn-datasetin kaikkien analogiakyselyiden oikeellisuusprosentti eri n-grammien pituuksilla. i on minimipituus, j maksimi.

$i\backslash j$	2	3	4	5	6	7	8	9	10
2	**44.74**	39.64	36.51	37.99	40.95	44.08	44.74	44.57	46.55
3		**41.61**	37.17	39.97	38.49	42.11	43.91	45.23	39.97
4			**39.97**	39.80	37.34	40.95	43.59	43.09	44.41
5				**40.30**	38.49	40.30	41.78	42.76	42.76
6					**41.28**	41.61	46.55	48.36	49.67
7						**45.56**	50.99	51.97	51.64
8							**52.30**	**56.91**	53.45
9								52.63	51.32
10									**56.41**

on 30,95%, mikä on sarakkeen suurin arvo.

Kun tarkastellaan kaikenlaisia kyselyitä SSWR-fi-datasetillä, huomataan, ettei n-grammien pituudella ole juurikaan väliä lopputuloksen kannalta. Vaikka taulukossa on merkitty kaikki mahdolliset yhdistelmät 2-grammeista 10-grammeihin saakka, ei taulukossa esiintyvä vaihtelu ole suurta. Paras valinta vaikuttaisi olevan 3-8-grammit, joilla oikeellisuusprosentti on 32,04%, mutta se ei ole merkittävästi parempi kuin kokeiden huonoin yhdistelmä, 7-8-grammit, joilla oikeellisuusprosentti on 28,02%.

Suurempia eroja löytyy kuitenkin, kun tarkastellaan merkitys- ja muotokyselyitä erikseen. Merkityskyselyistä nähdään, että sarakkeen paras arvo asettuu kaikissa kokeilluissa tapauksissa sarakkeen alareunaan. Se tarkoittaa, että jos suurimmaksi mukaan otetuksi n-grammiksi määrätään esimerkiksi 7, parhaan tuloksen saamiseksi kannattaa valita myös n-grammien alarajaksi 7, eli käytössä on vain 7-grammit, eikä mitään muita n.grammeja. Tällä valinnalla mallilla on mahdollisimman vähän n-grammeihin perustuvaa tietoa käytössään. Samoin taulukosta huomataan, että mitä suurempi n-grammi on, sitä paremmin malli on vastannut kyselyihin. Myös pelkästään suuria n-grammeja käyttämällä voidaan vähentää n-grammeihin perustavaa tietoa, sillä sanaan mahtuu enemmän lyhyitä kuin pitkiä n-grammeja. Monessa sanaston sanassa ei ole yhtäkään 10-grammia, jolloin 10-grammien käyttö ei anna ollenkaan lisää informaatiota.

Merkitystä koskevien kyselyiden taulukko on siis linjassa edellisen luvun tulosten kanssa, joissa havaittiin, että alimerkkijonoja hyödyntämättömät mallit pärjäsivät paremmin kuin mallit, jotka käyttivät n-grammitietoa osana vektoriesitystä.

Muotoa koskevilla kyselyillä tilanne on lähes päinvastainen kuin merkitystä koskevilla. Riippumatta siitä, mikä on pisin käytössä oleva n-grammi, on taulukon parhaat tulokset saatu lähes aina valitsemalla n-grammin minimiksi 3. Ainoastaan kun maksimipituus on 2 tai 3, parempi minimi on 2. Kun katsellaan toiseksi ja kolmanneksi parhaita vaihtoehtoja minimi-n-

Taulukko 9: SSWR-fi-datasetin top3-analogiakyselyiden oikeellisuusprosentti eri n-grammien pituuksilla. Tässä analogiakyselyt on toteutettu niin, että riittää, että kyselyn oikea vastaus löytyy kolmen parhaan ehdotuksen joukosta.

(a) Kaikki kyselyt.

$i \backslash j$	2	3	4	5	6	7	8	9	10
2	**39,34**	37,84	41,86	44,51	**45,73**	**47,10**	**47,48**	**47,24**	**48,08**
3		**41,13**	42,40	44,40	**45,73**	47,06	47,21	46,90	47,23
4			**43,88**	**45,18**	45,43	45,69	45,96	46,37	47,28
5				45,13	44,56	44,79	45,41	46,19	45,72
6					44,02	44,28	44,60	44,19	45,47
7						42,22	41,90	43,54	43,52
8							42,94	42,54	43,18
9								43,26	43,53
10									44,04

(b) Merkitystä koskevat kyselyt.

$i \backslash j$	2	3	4	5	6	7	8	9	10
2	**27,11**	19,15	23,89	28,32	30,24	33,33	34,27	33,96	35,03
3		**23,93**	24,00	27,28	29,97	32,08	32,03	32,54	32,68
4			**27,14**	29,35	30,09	31,31	32,26	32,54	34,79
5				**31,51**	31,81	31,96	33,44	34,80	34,93
6					**33,96**	34,50	34,99	35,72	37,35
7						**34,60**	34,79	37,28	37,74
8							**38,34**	38,10	39,40
9								**40,76**	41,08
10									**42,07**

(c) Muotoa koskevat kyselyt.

$i \backslash j$	2	3	4	5	6	7	8	9	10
2	**63,85**	75,31	77,90	76,94	76,78	74,71	73,97	73,86	74,24
3		**75,61**	**79,29**	**78,72**	**77,33**	**77,08**	**77,65**	**75,69**	**76,40**
4			77,44	76,92	76,18	74,52	73,42	74,08	72,31
5				72,44	70,12	70,50	69,39	69,00	67,37
6					64,20	63,87	63,85	61,17	61,75
7						57,49	56,15	56,10	55,09
8							52,17	51,43	50,75
9								48,27	48,46
10									42,07

grammeille maksimi-n-grammien mukaan, huomataan, että myös 2-grammit ja 4-grammit toimivat usein hyvin, ja minimiä kasvatettaessa laatu huonontuu. Siis mallit selvästi hyötyvät alimerkkijonojen sisältämästä informaatiosta, koska paljon sellaista informaatiota sisältävät mallit toimivat muotokyselyissä parhaiten. Kaikista paras n-grammivalinta näyttäisi taulukon perusteella olevan 3-4-grammien käyttäminen. Lähes yhtä hyvin toimivat kuitenkin myös esimerkiksi 2-4-grammit ja 3-5-grammit.

Myös muotokyselyitä koskevat tulokset siis tukevat edellisessä luvussa havaittua ilmiötä, jossa n-grammitieto edistää johdannaisia ja sananmuotoja koskevien analogioiden onnistumista. Erityisesti pienistä n-grammeista näyttää olevan apua.

Tarkastellaan seuraavaksi FinnAn-datasettiä koskevia tuloksia eri n-grammien pituuksien soveltamisesta. Myös nämä kyselyt ovat top1-kyselyitä. Tulokset on esitetty taulukossa 8. Havaitaan, että FinnAn-datasetillä tulokset muistuttavat paljon SSWR-fi-datasetin merkityskyselyiden tuloksia: yleisesti ottaen mitä suurempi minimi on valittu kutakin maksimi-n-grammia kohden, sitä paremmin malli on pärjännyt kyselyissä. Lisäksi kun mukaan on otettu vain suuria n-grammeja, kyselyiden oikeellisuus on parantunut. Siis mitä vähemmän n-grammitietoa mallille on annettu, sitä suurempi onnistumisprosentti on ollut. Tulos on linjassa aiemman kanssa, sillä FinnAn-datasetti koostuu lähes pelkästään merkityskyselyistä.

Tarkastellaan vielä SSWR-fi-datasettiä niin, että kokeillut analogiakyselyt ovatkin top3-muotoisia top1-kyselyiden sijasta. Tulokset on raportoitu taulukoissa 9a - 9c. Kautta linjan kyselyiden tulokset ovat linjassa top1-kyselyiden kanssa. Kun tarkastellaan molempia kyselytyyppejä yhdessä, on eri n-grammiyhdistelmien välillä hiukan enemmän eroja kuin top1-kyselyiden kohdalla. Taulukon parhaat tulokset ovat sellaisia, jossa maksimi-n-grammi on melko suuri ja minimi-n-grammi melko pieni. Paras tulos on saatu valitsemalla 2-10-grammit ja huonoin 2-3-grammeilla.

Merkitystä koskevat top3-kyselyt käyttäytyvät samalla tavalla kuin top1-kyselyt. Paras tulos on saatu 10-10-grammeilla ja huonoin 2-3-grammeilla. Mitä vähemmän n-grammitietoa mallilla on käytössä, sitä parempia tuloksia se on antanut. Myös muotoa koskevat top3-kyselyt ovat linjassa top1-kyselyiden kanssa. Sananmuotoja ja johdannaisia tutkittaessa malli on hyötynyt n-grammitiedoista. Paras tulos on jälleen saatu 2-4-grammeilla ja huonoin 10-10-grammeilla.

Tässä luvussa vahvistettiin edellisessä luvussa tehty havainto siitä, että n-grammitieto onnistumista parantaa muotokyselyiden ja huonontaa sitä merkityskyselyiden kohdalla. Seuraavassa luvussa pohditaan, mistä ilmiö voisi johtua, ja katsotaan, onko havainto linjassa aiemman tutkimuksen kanssa.

4.2.3 Tulosten vertailua

Bojanowski ym. [4] havaitsivat, että kun malli hyödyntää alimerkkijonotietoa, saksan- ja italiankielisillä aineistoilla muotokyselyiden laatu parani selvästi, mutta merkityskyselyiden laatu laski hieman. Italia ja saksa ovat suomen lailla kieliä, joissa sanojen taivuttamisella on suuri merkitys, joten tulos on linjassa tämän tutkielman kanssa. Myös Bojanowski ym. tutkivat n-grammien valinnan vaikutusta kyselyissä onnistumiseen, mutta he kokeilivat n-grammeja ainoastaan väliltä $[2, 6]$ eli ei $[2, 10]$, kuten tässä tutkielmassa.

Bojanowskin ym. tekemissä englantia ja saksaa koskevissa merkityskyselykokeissa saatiin samansuuntaisia tuloksia kuin tässä tutkielmassa taulukoissa 7b ja 9b: heillä paras valinta oli ottaa esitykseen mukaan 6-6-grammit (englannissa yhtä hyvin toimivat myös 5-5-grammit), ja jokaista maksimi-n-grammia kohtaan paras minimi oli maksimi itse, eli muotokyselyissä n-grammitietoa kannatti soveltaa mahdollisimman vähän. Ero korostui saksankielisellä aineistolla, mutta myös englanninkielisellä aineistolla tämä piti paikkansa.

Muotokyselyissä Bojanowski ym. havaitsivat, että parhaat tulokset saatiin sisällyttämällä malliin myös lyhyempiä n-grammeja. Esimerkiksi saksan kohdalla toimivat hyvin niin 3-6-grammit, 4-5-grammit, 4-6-grammit, 5-5-grammit kuin 5-6-grammitkin, englannin kohdalla parhaat tulokset antoivat 3-5-grammit, 4-5-grammit ja 4-6-grammit. Tulosten perusteella on siis mahdollista, että suomen kohdalla lyhyempien n-grammien merkitys muotokyselyiden kohdalla on korostunut saksaan ja englantiin verrattuna, sillä tutkielmaa varten tehtyjen kokeiden perusteella 3-grammien sisällyttäminen on aina kannattavampaa kuin niiden pois jättäminen.

Bojanowski ym. esittivät, että merkityskyselyissä tapahtunut huonontuminen on kompensoitavissa optimaalisella n-grammivalinnalla. Heidän englannilla ja saksalla tekemissään kokeissa havaitaankin, että kun n-grammit valitaan sopivasti, merkityskyselyiden laatu ei putoa alimerkkijonoja hyödyntämättömään malliin verrattuna, mutta muotokyselyiden laatu paranee, vaikkakaan paraneminen ei ole yhtä suurta kuin muotokyselyiden kannalta optimaalisesti valituilla n-grammeilla.

Tutkielmaa varten tehtyjen kokeiden perusteella mikään n-grammeja hyödyntävä suomenkielinen malli ei pääse merkityskyselyissä aivan alimerkkijonoja käyttämättömän mallin tasolle: kun alimerkkijonot eivät ole käytössä ja $d = 300$, on merkityskyselyiden onnistumistarkkuus 28,92%, kun taas alimerkkijonojen kanssa parhaimmillaankin eli 10-10-grammeilla 27,74%. Toisaalta 10-10-grammeilla muotokyselyiden onnistumistarkkuus on 35,50%, eli sekin on hiukan huonompi kuin ilman alimerkkijonoja, jolloin tarkkuus on 36,26%. Suomen kielelle on siis ilmeisesti vaikea löytää sellaista n-grammiväliä, jolla merkityskyselyiden tulos pysyisi lähes yhtä hyvänä kuin ilman n-grammeja hyvänä, mutta muotokyselyiden tulos paranisi n-grammien käytön ansiosta.

Jos halutaan optimoida molempien kyselyiden laatua, yksi vaihtoehto on ottaa koko datasettiä ajatellen paras n-grammiväli eli 3-8-grammit, joilla

kaikkien SSWR-fi-kyselyiden onnistumistarkkuus on 32,04%. Huomataan kuitenkin, että iso osa hyvästä tarkkuudesta seuraa nimenomaan onnistumisesta muotokyselyissä (62,35%, kun ilman alimerkkijonoja olisi 36,26%) ja merkityskyselyt ovat jo kärsineet reilusti (16,92%, kun ilman alimerkkijonoja olisi 28,92%). Suomen kielelle ei siis vaikuta pätevän, että optimaalisella n-grammivalinnalla voidaan saada malli, joka pärjää hyvin sekä muoto- että merkityskyselyissä. Toisaalta koko datasetin onnistumistarkkuutta katsellessa n-grammit eivät merkittävästi huononna analogiatulosten laatua: ilman alimerkkijonoja tarkkuus on 31,27%, kun $d = 300$, mikä on hiukan vähemmän kuin optimaalisin valinta 32,04% eikä edelleenkään merkittävästi enemmän kuin kokonaisuuden kannalta epäoptimaalisin valinta eli 27,98% (7-7-grammeilla).

Kaiken kaikkiaan paras kokeista löytynyt malli koko SSWR-fi-datasetille oli top1-kyselyillä $d = 100$ ja 4-7-grammit, jolla tulos oli 36,24% ja top3-kyselyillä $d = 200$ ja 3-6-grammit (50,65%). Huomataan siis, että kun sekä ulottuvuuksista että alimerkkijonoista kokeillaan eri yhdistelmiä, alimerkkijonojen muotokyselyissä tuoma hyöty nostaa niiden käytön tällä datasetillä alimerkkijonoja käyttämättömien mallien ohi.

Merkityskyselyissä alimerkkijonojen käyttö ei kannata. Koko SSWR-fi-datasetin paras tulos saatiin valitsemalla $d = 200$ ilman alimerkkijonoja (top1-kyselyillä 31,11% ja top3-kyselyillä 45,01%). Merkityskyselypainotteisella FinnAn-datasetillä tulos oli samankaltainen: 8-9-grammeilla ($d = 300$) saatu kaikista kokeista paras tulos 56,91% ei ole juurikaan parempi kuin ilman alimerkkijonoja ($d = 300$) saatu 55,92%.

Muotokyselyissä SSWR-fi-datasetillä kaiken kaikkiaan paras tulos saatiin valitsemalla $d = 300$ ja 3-4-grammit (top1-kyselyissä 65,16% ja top3-kyselyissä 79,29%).

Optimaalisista valinnoista huolimatta suomenkielisellä SSWR-fi-datasetillä ei päästä vielä lähelle englanninkielisen datasetin tarkkuutta, joka on Bojanowski ym [4] mukaan 3-6-grammeilla ja $d = 300$ merkityskyselyillä 77,8% ja muotokyselyillä 74,9%.

4.2.4 Mistä ero muoto- ja merkityskyselyiden välillä voisi johtua?

Ajatus n-grammien käyttämisestä sanojen vektoriesityksen luomisen apuna perustuu sanojen taivutusmuotojen huomioimiseen. Sellaisissa kielissä, joissa sanoja taivutetaan ja johdetaan, sanat eivät ole toisistaan erillisiä, vaan niitä muodostetaan toistensa pohjalta. Siksi on myös hyvä, jos malli voi tunnistaa, että sanoissa on samoja alimerkkijonoja. Jos sana on taivutusmuoto, johdannainen tai yhdyssana, alimerkkijonot kertovat sanojen välisestä yhteydestä. Esimerkiksi sanoilla *susi* ja *suden* on yhteinen alkuosa *su*, koska ne ovat saman sanan taivutusmuotoja, ja sanoilla *suden* ja *käden* on yhteinen loppuosa *den*, koska ne ovat samassa taivutusmuodossa.

Yksi mahdollinen selitys muotokyselyiden onnistumiselle on se, että mo-

net aineiston johdannaiset ja sananmuodot eroavat sanan perusmuodosta vain muutaman merkin pituisella morfeemilla eli minimaalisella merkitystä kantavalla yksiköllä. Esimerkiksi yksi SSWR-fi-datasetissäkin ilmenevä tapa muodostaa suomen kielen vastakohtia on lisätä sanan alkuun *epä*, kuten sanaparissa *selvä, epäselvä*. Tässä *epä* on lyhyt morfeemi: koska sanan alkua merkitään <-merkillä, kuuluu se alimerkkijonoon *<epä*, jonka pituus on 4 merkkiä. Kun tällaiset lyhyet alimerkkijonot otetaan mukaan sanan vektoriesitystä muodostaessa, malli oppii sanat osittain niiden perusteella. Kun tutkitaan erilaisia n-grammien välejä, havaitaan, että erityisesti lyhyiden n-grammien ottaminen mukaan vaikuttaisi parantavan muotokyselyiden laatua ja vähentävän merkityskyselyissä onnistumista. Paras valittu väli suomelle vaikutti olevan 3-4-grammit, ja 3-grammien jättäminen pois huononsi aina muotokyselyissä onnistumista.

Bojanowski ym. [4] mainitsevat, että englannin ja saksan kohdalla 2-grammeista ei ole juuri hyötyä, koska aloitus- ja lopetusmerkit < ja > pidentävät prefiksejä ja suffikseja yhdellä merkillä, jolloin 2-grammista jää enää yksi merkki sanan osia merkitsemään. Suomen kielessä taivuttaminen ja johtaminen eivät aina toimi niin, että sanan alkuun tai loppuun lisättäisiin merkitysyksikkö, koska sanassa voi olla useita päätteitä, joista vain osaa muutetaan sanaa taivuttaessa. Esimerkiksi sanaparissa *kissamme, kissallamme* lisätään adessiivia ilmaiseva *lla* sanan perusmuodon ja omistusliitteen *mme* väliin. Testiaineistoissa sanat ovat kuitenkin muotoa, jossa sanan alusta tai lopusta korvataan joitakin osia toisilla taivutuksen tuottamiseksi. Testiaineistossa ei ole myöskään tapauksia, joissa taivutuksia olisi suomelle ominaiseen tapaan yhdistelty, vaan kyse on aina perusmuodon ja yhdentyyppisen taivutuksen tai johdannaisen vertailupareista. Toisaalta taivutusta ilmaisevat osat voivat olla melko lyhyitä, kuten sanaparin *nuori, nuorin* superlatiivia ilmaiseva *n*.

Toisin kuin muotokyselyissä, merkityskyselyissä sanan muodosta ei voi välttämättä päätellä mitään siitä, minkälainen suhde sanoilla on. Ei ole väliä, että sanoissa *Intia* ja *rupia* esiintyy yhteinen alimerkkijono *ia* ja sanoissa *Brasilia* ja *real* ei, sillä sanojen suhde **maa ja valuutta** ei riipu siitä, mistä merkeistä sanat koostuvat, vaan ainoastaan kontekstista. Siksi alimerkkijonojen korostaminen voi hämätä mallia painottamalla sanojen sisältämiä merkkejä, vaikka näissä kyselyissä ainoastaan sanojen kontekstilla on väliä.

4.2.5 Havaintoja yleisistä virheistä

On kiinnostavaa tutkia, millaisia virheitä malli tekee. Virheitä tarkastelemalla voidaan saada tietoa siitä, miten mallia pitäisi muuttaa, jotta se toimisi useammin oikein.

Taulukossa 10 nähdään esimerkkejä neljästä eri virhetyypistä. Ylimpänä on vääriä vastauksia, kuten *emotionaalinen* sanan *irrationaalinen* sijasta,

Kysely			Oikea vastaus	Parhaat tulokset
epälooginen -	looginen +	rationaalinen	irrationaalinen	emotionaalinen, epätodellinen, rationaalisuus
paras -	hyvä +	nopea	nopein	maalintekijä, tehokkain, nopein
ui -	uida +	ajatella	ajattelee	ajatellaan, formaalisti, järjestelmäviestistä
mahtavasti -	mahtava +	ilmeinen	ilmeisesti	ilmeisen, ilmeisestikin, ilmeisin
mahtavasti -	mahtava +	hilpeä	hilpeästi	hilpeää, hilpeästi, hilpeän
epäeettinen -	eettinen +	mukava	epämukava	mukavaa, mukavan, mukavampi
mahtavasti -	mahtava +	täydellinen	täydellisesti	mahtavan, mahtavaa, mahtavalta
mahtavasti -	mahtava +	onnekas	onnekkaasti	mahtavalta, mahtavin, mahtavan
vapaasti -	vapaa +	iloinen	iloisesti	vapaata, vapaastikantava, vapa
hiljaisesti -	hiljainen +	tarkka	tarkasti	tarkkaavaisesti, tarkkaan, tarkasti
heikoin -	heikko +	hyvä	paras	parhain, hyvähän, erinomaisin
käärmeet -	käärme +	vuohi	vuohet	vuohieläimet, vuohiset, vuohieläin

Taulukko 10: Vääriä tuloksia muotokyselyihin mahdollisimman hyvin toimivalla yhdistelmällä eli $d = 300$, 3-4-grammit. Taulukossa on esimerkkejä neljästä erityyppisestä virheestä: selvästi väärät vastaukset, oikean vastauksen väärät muodot, kyselyn sanojen väärät muodot ja tavallaan oikeat, mutta ei haetut vastaukset.

sillä sanat eivät tarkoita samaa, vaikka esiintyvätkin usein samankaltaisissa yhteyksissä. Toiseksi ylimpänä on vääriä sananmuotoja haetusta sanasta, esimerkiksi *hilpeää* muodon *hilpeästi*. Väärä muoto voi aiheuttaa sen, että kyselyn vastaus on oleellisesti väärä: esimerkiksi *mukavaa* on **vastakohta**-kyselyssä aivan eri asia kuin haettu *epämukava*. Toiseksi alimpana on vääriä sananmuotoja kyselyssä esiintyneistä sanoista, esimerkiksi sanalla *vapaasti* alkavan kyselyn tulos *vapaata*. Välillä sana ei ole oikeasti sananmuoto kyseisestä sanasta, sillä *vapa* ei ole sanan *vapaa* muoto, vaikka niillä onkin pitkä yhteinen alimerkkijono. Alimpana on vastauksia, jotka ovat tietyssä mielessä oikein, mutta eivät kuitenkaan kyselyssä haettuja, esimerkiksi *vuohieläimet*, kun pitäisi olla *vuohet*.

Taulukostakin siis havaitaan, että usein mallin tarjoama väärä vastaus oli jokin väärä sananmuoto joko haetusta sanasta w_4 tai sitten joistakin kyselyssä esiintyneistä sanoista w_1, w_2 ja w_3. Tehtävänanto on määritelty testissä niin, että jos mallin tarjoama vastaus esiintyy kyselyssä (eli on w_1, w_2 tai w_3), hylätään vastaus ja katsotaan seuraavaksi tarjottua sanaa. Vähän taivuttavan englannin kohdalla tämä päätös on yksinkertaisempi kuin suomen kohdalla, sillä suomessa voitaisiin argumentoida myös, että esimerkiksi kaikki kyselyn sanojen sananmuodot hylätään vastausten joukosta. Nyt esim. kyselyllä *Helsinki - Suomi + Saksa* hylättiin ensimmäisenä tarjottu vastaus *Helsinki*, mutta ei vastausta *Helsingin*, mikä teki suomen kielellä kyselyistä siinä mielessä haastavamman, että oikealle vastaukselle ylipäätään löytyy melkein oikea mutta väärässä muodossa oleva vastaus. Englannissa ongelmaa ei samassa mittakaavassa ole.

Erityisesti merkityskyselyissä voidaan kyseenalaistaa se, onko oikean vastauksen pakko olla samassa sananmuodossa kuin kyselyn sanat. Muotokyselyissä w_4:n väärien muotojen hyväksyminen ei luonnollisestikaan olisi perusteltua, mutta edelleen jää kysymys, pitäisikö muiden w_i taipuneet muodot hylätä, kun kyselyssä esiintyneetkin sanat hylätään.

Tällä tavalla meneteltäessä on kuitenkin myös vaikea sanoa, mihin raja tulisi vetää. Toinen vaihtoehto olisikin suorittaa kyselyt niin, että vaikeutetaan tehtävää ja jätetään hylkäämättä mitään sanoja, vaikka ne esiintyisivät kyselyssä. Tämä voisi mahdollistaa paremman vertailun suomen ja englannin välillä, mutta se vaatisi samalla tavalla toteutetun arvioinnin tekemistä myös englannille. Tässä tutkielmassa suoritettiin kyselyt vertailukelpoisuuden vuoksi samalla tavalla kuin aiemmassa kirjallisuudessa, vaikka menetelmän ongelmat olivat tiedossa.

Voidaan myös kyseenalaistaa se, onko oikean vastineen löytyminen top1- tai top3-joukosta paras mahdollinen mittari mallille. Sopiva metriikka arvioimaan mallin kykyä vastata kyselyihin voisi olla esimerkiksi käänteinen järjestys (inverse rank), jossa ensimmäinen vaihtoehto saa 1 pisteen, toinen $\frac{1}{2}$ pistettä, kolmas $\frac{1}{3}$ pistettä ja niin edelleen.

5 Johtopäätökset ja tuleva tutkimus

Tässä tutkielmassa selvitettiin, parantaako alimerkkijonojen käyttäminen suomen kielen sanojen vektoriesitysten laatua. Tätä tutkittiin testaamalla erilaisilla n-grammien pituuksilla ja sanavektorin ulottuvuuksilla opetettuja fastText-malleja. Testit arvioivat mallin kykyä suoriutua semanttista samankaltaisuutta ja sana-analogioiden oikeellisuutta mittaavista tehtävistä.

Semanttista samankaltaisuutta tutkittaessa havaittiin, että alimerkkijonojen käyttö parantaa mallin laatua. Analogiatesteissä huomattiin, että alimerkkijonojen käyttö parantaa mallin kykyä suoriutua muotoon liittyvistä kyselyistä, mutta huonontaa sen selviytymistä merkitykseen liittyvien kyselyiden kohdalla.

5.1 Onko alimerkkijonojen käyttö suomen kielellä hyödyllistä?

Tässä tutkielmassa haluttiin selvittää, pitääkö Bojanowskin ym. [4] esittämä hypoteesi alimerkkijonojen hyödyllisyydestä suomen kielellä paikkansa. Bojanowski ym. esittivät, että aiempien menetelmien luomat mallit, jotka esittävät jokaisen sanan erillisenä merkkijonoja, ovat merkittävästi rajoittuneita vahvasti taivutukseen nojaavien kielten kuin suomen kohdalla. Merkkitasoa käyttävillä menetelmillä voidaan parantaa tällaisten kielten malleja.

Tutkielmassa tehdyt havainnot puoltavat Bojanowskin ym. hypoteesia. Erityisesti semanttista samankaltaisuutta mittaavissa testeissä tulokset paranivat, kun hyödynnettiin alimerkkijonotietoa. Ainoastaan SimLex-datasetillä ja $d \leq 300$ alimerkkijonojen käyttö ei juuri vaikuttanut mallin oppimiseen kokeilluilla 3-6-grammeilla ja 4-7-grammeilla. Kuitenkin silläkin datasetillä alimerkkijonotieto alkoi parantaa mallin laatua, kun d oli tarpeeksi suuri. Erot alimerkkijonojen kanssa ja ilman näkyivät paremmin WS353- ja WS277-dataseteillä, jotka mittaavat samaan aihepiiriin kuulumista, mikä on helpompi tehtävä. Kuitenkin myös samaa merkitystä mittaavilla SimLex- ja FinnSim-dataseteillä alimerkkijonojen käytöstä oli hyötyä.

Epäoptimaalinen n-grammivalinta tuotti kuitenkin huonompia tuloksia kuin alimerkkijonojen pois jättäminen. Kokeissa havaittiin, että jos mallin opetukseen otetaan mukaan esimerkiksi vain 2-2-grammit, on sen laatu huonompi kuin ilman alimerkkijonoja semanttista samankaltaisuutta mitattaessa. Alimerkkijonojen käyttö on hyödyllistä vain, jos merkkijonot valitaan edes jotenkuten optimaalisesti, mikä tahansa valinta ei kelpaa.

Sana-analogioita mittaavissa testeissä näyttää kaiken tyylisiä analogioita tarkastellessa siltä kuin alimerkkijonojen hyödyntämisellä ei olisi suurta vaikutusta. Kun tarkastellaan merkitystä ja muotoa mittaavia kyselyitä erikseen, huomataan kuitenkin, että alimerkkijonojen käyttäminen heikentää merkityskyselyiden laatua ja parantaa samanaikaisesti muotokyselyiden laatua. Tämä on siinä mielessä varsin intuitiivinen tulos, että juuri muotokyselyiden

eri muodot koostuvat merkkijonoista. Kaikki sananmuodostus ei tapahdu samalla tavalla, koska suomen kielen sanamuotoja muodostettaessa ei riitä vain lisätä sanan perään pääte eli suffiksi, vaan taivuttaminen vaikuttaa myös sanan vartaloon. Kuitenkin taivutuksissa on tiettyjä sääntöjä, joita malli oppii paremmin, jos sille annetaan tietoa alimerkkijonoista. Sen sijaan merkityskyselyissä alimerkkijonojen käyttö tuntuu hämäävän mallia, koska kyselyt ovat luonteeltaan sellaisia, ettei sanan sisältämistä merkeistä voi päätellä oikeaa vastausta. Lisäksi yksi virheiden lähde oli se, että mallin tarjoama vastaus oli joku haetun sanan tai kyselyssä esiintyneen sanan muu sananmuoto. Tämä vaikeuttaa myös vertailua englannin kielelle saatuihin tuloksiin kirjallisuudessa, koska englannissa vastaavaa ongelmaa ei ole yhtä laajana.

Vaikka alimerkkijonotieto paransi mallin laatua, ei optimaalisinkaan tutkielmassa tarkasteltu malli päässyt lähelle englanninkielisten mallien tuloksia samoilla käännetyillä dataseteillä. Osa eroista selittyy varmasti käännöksessä menetetyn informaation vaikutuksesta, sillä esimerkiksi homonymia, synonymia ja polysemia eivät säily käännöksessä. Voidaan kuitenkin myös sanoa, että alimerkkijonojen käyttö ei riitä tasoittamaan eroa vain vähän taivuttavan englannin ja vahvasti sananmuotoihin ja sananmuodostukseen nojaavan suomen välillä.

Toinen mahdollinen syy suomen ja englannin tulosten välillä on opetusaineiston koko. Suomenkielinen Wikipedia, jota tutkielmassa käytettiin opetusaineistona, on paljon pienempi kuin esim. Bojanowskin ym. [4] hyödyntämä englanninkielinen Wikipedia. Kuitenkin Bojanowski ym. havaitsivat myös, että fastText-mallien laatu ei juuri huonontunut, vaikka käytettäisiin vain 1% opetusaineistosta. Suurempi opetusaineisto ei siis välttämättä lisää mallien laatua.

5.2 Mitkä parametrit toimivat parhaiten?

Tutkielmassa ei löytynyt yksikäsitteisesti parhaita parametreja kaikkiin suoritettaviin tehtäviin. Tarkastellut parametrit olivat sanavektorin ulottuvuus d sekä mukaan vektoriesitykseen otetut n-grammit.

Semanttista samankaltaisuutta mittaavissa testeissä parempia tuloksia antoivat suuremmat d:n arvot kuin analogiakyselyissä. Semanttista samankaltaisuutta mittaavissa testeissä $d = 500$ vaikutti parhaalta arvolta. Myös $d = 300$ toimi hyvin, mutta $d = 200$ tai vähemmän oli jo selvästi huonompi. Analogiatesteissä taas $d \in \{100, 200, 300\}$ toimivat hyvin, mutta sitä suuremmilla arvoilla tulos alkoi huonontua. On siis vaikea löytää sellaista d:n arvoa, joka olisi kummallakin testityypillä paras, mutta monissa tapauksissa $d = 300$ vaikuttaisi toimivan melko hyvin.

Tätäkin vaikeampaa on valita parhaat n-grammit, jos halutaan menestyä kummassakin tehtävätyypissä. Semanttista samankaltaisuutta mittaavissa testeissä isompien n-grammien mukaan ottaminen vaikutti hyödylliseltä:

hyviä tuloksia saatiin esimerkiksi 4-9-grammeilla, 3-10-grammeilla ja 5-8-grammeilla. Tosin SimLex-datasetillä myös pelkät 6-6-grammit toimivat erinomaisesti. Usein pienillä n-grammeilla ei ollut juuri väliä, vaan niiden mukaan ottaminen laski mallin laatua hieman.

Analogiatesteissä havaittiin, että n-grammien mukaan ottaminen paransi muotokyselyiden ja laski merkityskyselyiden laatua. Vaikutti mahdottomalta löytää sellainen n-grammiväli, jolla molemmat kyselytyypit olisivat toimineet hyvin. Kun tarkastellaan kaikkia kyselyitä, näyttikin siltä kuin n-grammien valinnalla ei olisi juuri väliä. Tämä johtui kuitenkin siitä, että merkityskyselyiden oikeellisuus huonontui muotokyselyiden oikeellisuuden kasvaessa, jolloin keskimääräinen oikeellisuus ei parantunut tai huonontunut merkittävästi.

Toisin kuin semanttisen samankaltaisuuden kohdalla, analogiatesteissä isoista n-grammeista ei ollut merkittävää hyötyä edes muotokyselyissä, vaan niiden mukaan ottaminen laski mallin laatua hieman. Vielä isompaa laskua aiheutti pienten n-grammien pois jättäminen. Paras muotokyselyiden n-grammiväli oli 3-4-grammit, mutta myös sitä lähellä olevat välit toimivat hyvin. Merkityskyselyissä paras tulos saatiin, kun n-grammitietoa ei käytetty ollenkaan, ja siksi parhaita välejä tutkittaessa 10-10-grammit toimivat parhaiten, koska niissä ei yksinään ole juurikaan lisätietoa.

Riippuu siis tehtävästä, millaiset parametrit kannattaa valita. Bojanowskin ym. [4] oletusparametrit eli $d = 300$ ja 3-6-grammit toimivat myös suomen kielellä kohtalaisesti. Monissa käytännön tehtävissä samankaltaisten sanojen lähekkäisyys on tärkeämpää kuin oikeiden analogiasuhteiden löytyminen sana-avaruudesta. Siis esimerkkejä hyviin parametrivalintoihin käytännön tehtäviin ovat $d = 500$ ja 4-7-grammit tai $d = 300$ ja 5-9-grammit. On vaikea sanoa, mitkä parametrit toimivat kaikissa käytännön tapauksissa parhaiten, vaan sopivia parametreja joutuu aina kokeilemaan. Tutkielmassa tehtyjen kokeiden perusteella voidaan kuitenkin rajata joitakin parametreja pois todennäköisesti hyvin toimivien joukosta. Esimerkiksi suuret d:n arvot ($d > 500$) voi mahdollisesti jättää kokeilematta, sillä tätä pienemmät arvot näyttäisivät toimivan paremmin, ja d:n kasvattaminen hidastaa verkon opetusta.

5.3 Tuleva tutkimus

Tässä tutkielmassa kokeiltiin laajasti eri vektorin ulottuvuuksia ja alimerkkijonovälejä fastText-skip-gram-mallille. Aikaisemmassa tutkimuksessa on joskus havaittu, että CBOW-malli toimii skip-gram-mallia paremmin eri tilanteissa [42]. Tutkielmassa ei onnistuttu toistamaan tätä tulosta. Voisi kuitenkin olla hyödyllistä tutkia, olisiko joissakin tapauksissa CBOW-malli skip-gram-mallia parempi tutkielman käsittelemissä testityypeissä.

Semanttinen samankaltaisuus ja analogiatestit eivät ole ainoita mahdollisuuksia sanavektorien laadun teoreettiseen tutkimiseen. Voisi olla kiinnostavaa nähdä, miten alimerkkijonot vaikuttavat toisenlaisissa kokeissa. Yksi

mahdollinen koe on ihmisten älykkyyttä mitattaessakin käytetty intruusiotesti, jossa esimerkiksi viiden sanan joukosta täytyy osata valita yksi, joka ei kuulu joukkoon. Eli esimerkiksi joukossa *kissa, kala, tiikeri, vasara, perhonen* tulisi tunnistaa *vasara*, sillä se ei ole eläin.

Toinen kiinnostava tutkimuskohde olisi alimerkkijonojen käytön testaaminen jossakin käytännön ongelmassa, johon sanavektoreita voidaan käyttää. Esimerkki tällaisesta ongelmasta voisi olla luokittelutehtävä kuten sävyn (sentiment) tunnistaminen. Tämä tutkimussuunta olisi siinä mielessä hyödyllinen, että vektoriesityksiä voidaan hyödyntää useissa käytännön tehtävissä, ja se onkin yleensä syy sille, miksi laadukkaita sanavektoreita ylipäätään halutaan voida muodostaa.

Olisi kuitenkin kiinnostavaa laajentaa tutkimusta myös teoreettiselta kannata erityisesti analogiatestien osalta. Suomen kielelle on helppo muodostaa lisää taivutusanalogioita, koska suomen kiellä taivutusmuotoja on paljon. Voisi olla kiinnostavaa tutkia, voidaanko parametreja vaihtelemalla havaita eroa tietyn taivutusmuodon ja tietyn semanttista merkitystä kantavan muodon muodostamisessa. Esimerkiksi sijaintia voidaan suomen kielessä ilmaista sekä inessiivillä (esim. *talossa, Helsingissä*) että adessiivilla (esim. *niityllä, Tampereella*). Tämä olisi sanamuodostukseen liittyvä analogiatapaus, jossa kuitenkin semanttisella informaatiolla on merkitystä, jotta sananmuodostus tapahtuu oikein: ei siis riitä laittaa sanoihin suffiksia *ssa* tai *ssä*, jotta voisi ilmaista sijaintia, vaikka se on yleisempi tapaus.

Analogiatesteissä havaittiin myös, että alimerkkijonoja hyödyntäessä usein mallin tarjoama mutta väärä vastaus oli jokin väärä sananmuoto halutusta sanasta tai jostakin kyselyssä esiintyvästä sanasta. Voisi siis olla järkevää määritellä tehtävä uudelleen esimerkiksi siten, että vastauksia ei koskaan hylätä, vaikka ne esiintyisivätkin kyselyssä. Tarkemmin näitä mahdollisuuksia on pohdittu luvussa 4.2.5.

Muita kiinnostavia jatkotutkimuksen aiheita voisivat olla myös Bojanowskin ym. [4] toteuttama kvalitatiivinen analyysi sanojen merkittävimmistä n-grammeista sekä heidän mainitsemansa yksinkertaiset muutokset malliin. Esimerkiksi suomen kielellä tiedetään, että taivutus tapahtuu pääosin suffiksien avulla, joten niiden painotus prefikseihin ja infikseihin verrattuna voisi olla erilainen. Tällä hetkellä malli painottaa kaikkia sanan n-grammeja saman verran.

Tämän tutkielman tekemisen aikana on ilmestynyt lukuisia uusia fastText-menetelmää hyödyntäviä tutkimuksia [12][7][41]. Alimerkkijonojen hyödyntäminen on siis kiinnostusta herättävä menetelmä sanavektorien parantamiseen ja sitä tulisi tutkia myös suomen kielellä laajemmin.

Lähteet

[1] Eneko Agirre, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius

Pasca ja Aitor Soroa: *A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches*. Teoksessa *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, Boulder, Colorado, USA*, sivut 19–27, 2009.

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent ja Christian Janvin: *A Neural Probabilistic Language Model*. Journal of Machine Learning Research, 3:1137–1155, 2003.

[3] Giacomo Berardi, Andrea Esuli ja Diego Marcheggiani: *Word Embeddings Go to Italy: A Comparison of Models and Training Datasets*. Teoksessa *Proceedings of the 6th Italian Information Retrieval Workshop, Cagliari, Italy*, 2015.

[4] Piotr Bojanowski, Edouard Grave, Armand Joulin ja Tomas Mikolov: *Enriching Word Vectors with Subword Information*. Transactions of the Association for Computational Linguistics, 5:135–146, 2017.

[5] Peter F. Brown, John Cocke, Stephen Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer ja Paul S. Roossin: *A Statistical Approach to Machine Translation*. Computational Linguistics, 16(2):79–85, 1990.

[6] Walter G. Charles: *Contextual correlates of meaning*. Applied Psycholinguistics, 21:505–524, 2000.

[7] Shamil Chollampatt ja Hwee Tou Ng: *A Multilayer Convolutional Encoder-Decoder Neural Network for Grammatical Error Correction*. Teoksessa *AAAI*. AAAI Press, 2018.

[8] Mathias Creutz, Krista Lagus ja Sami Virpioja: *Unsupervised Morphology Induction Using Morfessor*. Teoksessa A. Yli-Jyrä, L. Karttunen ja J. Karhumäki (toimittajat): *Finite-State Methods and Natural Language Processing (FSMNLP 2005)*, nide 4002 sarjassa *Lecture Notes in Computer Science*, sivut 300–301. Springer-Verlag Berlin Heidelberg, 2006.

[9] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer ja Richard Harshman: *Indexing by latent semantic analysis*. Journal of the American Society for Information Science, 41(6):391–407, 1990.

[10] Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman ja Eytan Ruppin: *Placing search in context: the concept revisited*. ACM Trans. Inf. Syst., 20(1):116–131, 2002.

[11] J. Firth: *A synopsis of linguistic theory 1930-1955.* Studies in Linguistic Analysis, Philological. Longman, 1957.

[12] Carlos-Emiliano González-Gallardo ja Juan-Manuel Torres-Moreno: *Sentence Boundary Detection for French with Subword-Level Information Vectors and Convolutional Neural Networks.* CoRR, abs/1802.04559, 2018.

[13] Ian Goodfellow, Yoshua Bengio ja Aaron Courville: *Deep Learning.* MIT Press, 2016.

[14] Roger Granada, Cássia Trojahn dos Santos ja Renata Vieira: *Comparing Semantic Relatedness between Word Pairs in Portuguese Using Wikipedia.* Teoksessa *Computational Processing of the Portuguese Language - 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil. Proceedings*, sivut 170–175, 2014.

[15] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin ja Tomas Mikolov: *Learning Word Vectors for 157 Languages.* Teoksessa *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.*, 2018.

[16] Iryna Gurevych: *Using the structure of a conceptual network in computing semantic relatedness.* Natural Language Processing–IJCNLP 2005, sivut 767–778, 2005.

[17] Felix Hill, Kyunghyun Cho, Sébastien Jean, Coline Devin ja Yoshua Bengio: *Embedding Word Similarity with Neural Machine Translation.* arXiv e-prints, abs/1412.6448, joulukuu 2014.

[18] Felix Hill, Roi Reichart ja Anna Korhonen: *Simlex-999: Evaluating semantic models with (genuine) similarity estimation.* Computational Linguistics, 2015.

[19] Colette Joubarne ja Diana Inkpen: *Comparison of Semantic Similarity for Different Languages Using the Google n-gram Corpus and Second-Order Co-occurrence Measures.* Teoksessa *Advances in Artificial Intelligence - 24th Canadian Conference on Artificial Intelligence, Canadian AI 2011, St. John's, Canada. Proceedings*, sivut 216–221, 2011.

[20] Fred Karlsson: *Yleinen kielitiede.* Yliopistopaino, 1994.

[21] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin ja Kilian Q. Weinberger: *From Word Embeddings To Document Distances.* Teoksessa *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France*, sivut 957–966, 2015.

[22] Maximilian Köper, Christian Scheible ja Sabine Schulte im Walde: *Multilingual Reliability and "Semantic"Structure of Continuous Word Spaces.* Teoksessa Matthew Purver, Mehrnoosh Sadrzadeh ja Matthew Stone (toimittajat): *IWCS*, sivut 40–45. The Association for Computer Linguistics, 2015.

[23] Ira Leviant ja Roi Reichart: *Separated by an Un-common Language: Towards Judgment Language Informed Vector Space Modeling.* arXiv e-prints, arxiv:1508.00106, 2015.

[24] Thang Luong, Richard Socher ja Christopher D. Manning: *Better Word Representations with Recursive Neural Networks for Morphology.* Teoksessa *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria*, sivut 104–113, 2013.

[25] William C. Mann: *An overview of the PENMAN text generation system.* Teoksessa *Proceedings of the National Conference on Artificial Intelligence*, sivut 261–265. AAAI, elokuu 1983.

[26] Tomas Mikolov, Kai Chen, Greg Corrado ja Jeffrey Dean: *Efficient estimation of word representations in vector space.* arXiv preprint arXiv:1301.3781, 2013.

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado ja Jeffrey Dean: *Distributed Representations of Words and Phrases and their Compositionality.* Teoksessa *NIPS*, sivut 3111–3119, 2013.

[28] George A Miller ja Walter G Charles: *Contextual correlates of semantic similarity.* Language & Cognitive Processes, 6(1):1–28, 1991.

[29] Alexander Panchenko, Dmitry Ustalov, Nikolay Arefyev, Denis Paperno, Natalia Konstantinova, Natalia V. Loukachevitch ja Chris Biemann: *Human and Machine Judgements for Russian Semantic Relatedness.* CoRR, abs/1708.09702, 2017.

[30] Tommi Pirinen: *Omorfi — Free and open source morphological lexical database for Finnish.* Linköping University Electronic Press, Linköpings universitet, 2015.

[31] Marten Postma ja Piek Vossen: *What implementation and translation teach us: the case of semantic similarity measures in wordnets.* Teoksessa *Proceedings of the Seventh Global Wordnet Conference, GWC 2014, Tartu, Estonia*, sivut 133–141, 2014.

[32] Alec Radford, Rafal Józefowicz ja Ilya Sutskever: *Learning to Generate Reviews and Discovering Sentiment.* CoRR, abs/1704.01444, 2017.

[33] Philip Resnik: *Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language.* Journal of Artificial Intelligence Research, 11(1), 1998.

[34] Herbert Rubenstein ja John B. Goodenough: *Contextual correlates of synonymy.* Commun. ACM, 8(10):627–633, 1965.

[35] Hasim Sak, Murat Saraclar ja Tunga Güngör: *Morphology-based and sub-word language modeling for Turkish speech recognition.* Teoksessa *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, Dallas, Texas, USA*, sivut 5402–5405, 2010.

[36] Hinrich Schütze: *Dimensions of Meaning.* Teoksessa *Proceedings Supercomputing '92, Minneapolis, MN, USA*, sivut 787–796, 1992.

[37] Bayan Abu Shawar ja Eric Atwell: *Chatbots: Are they Really Useful?* LDV Forum, 22(1):29–49, 2007.

[38] Rion Snow, Brendan O'Connor, Daniel Jurafsky ja Andrew Y. Ng: *Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks.* Teoksessa *EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, sivut 254–263, Morristown, NJ, USA, 2008. Association for Computational Linguistics.

[39] C. Spearman: *The Proof and Measurement of Association Between Two Things.* American Journal of Psychology, 15:88–103, 1904.

[40] Lukás Svoboda ja Tomás Brychcín: *New Word Analogy Corpus for Exploring Embeddings of Czech Words.* Teoksessa *Computational Linguistics and Intelligent Text Processing - 17th International Conference, CICLing 2016, Konya, Turkey, Revised Selected Papers, Part I*, sivut 103–114, 2016.

[41] Julien Tissier, Christophe Gravier ja Amaury Habrard: *Dict2vec : Learning Word Embeddings using Lexical Dictionaries.* Teoksessa *EMNLP*, sivut 254–263. Association for Computational Linguistics, 2017.

[42] Viljami Venekoski ja Jouko Vankka: *Finnish resources for evaluating language model semantics.* Teoksessa *Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, Gothenburg, Sweden*, sivut 231–236. Linköping University Electronic Press, Linköpings universitet, 2017.

[43] Oriol Vinyals, Alexander Toshev, Samy Bengio ja Dumitru Erhan: *Show and Tell: A Neural Image Caption Generator.* Teoksessa *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[44] Joseph Weizenbaum: *ELIZA a Computer Program for the Study of Natural Language Communication Between Man and Machine*. Commun. ACM, 9(1):36–45, tammikuu 1966, ISSN 0001-0782.

[45] Will Y. Zou, Richard Socher, Daniel M. Cer ja Christopher D. Manning: *Bilingual Word Embeddings for Phrase-Based Machine Translation.* Teoksessa *EMNLP*, sivut 1393–1398. ACL, 2013, ISBN 978-1-937284-97-8.